

The 13th Conference on Artificial Intelligence in Medicine (AIME'11)

*Probabilistic Problem Solving
in BioMedicine*

ProBioMed'11

**July 2, 2011
Bled, Slovenia**

Organisation

Chairs

| | |
|-----------------|--|
| Arjen Hommersom | Radboud University Nijmegen, The Netherlands |
| Peter Lucas | Radboud University Nijmegen, The Netherlands |

Program Committee

| | |
|-----------------------|---|
| Juan-Manuel Ahuactzin | ProBayes, France |
| Nicos Angelopoulos | Netherlands Cancer Institute, The Netherlands |
| Péter Antal | BME, Hungary |
| Alexandre Aussem | University of Lyon 1, France |
| Riccardo Bellazzi | University of Pavia, Italy |
| Concha Bielza | Technical University of Madrid, Spain |
| Hendrik Blockeel | KU Leuven, Belgium |
| Luis M. de Campos | University of Granada, Spain |
| Andrés Cano | University of Granada, Spain |
| Robert Castelo | Pompeu Fabra University, Spain |
| Li Cheng | A*STAR, Singapore |
| Jesse Davis | KU Leuven, Belgium |
| Javier Díez | UNED, Spain |
| Marek J Druzdzel | University of Pittsburgh, USA |
| | Bialystok University of Technology, Poland |
| Norman Fenton | Queen Mary University of London, UK |
| Julia Flores | UCLM, Spain |
| Linda van der Gaag | Utrecht University, The Netherlands |
| Jose A. Gámez | UCLM, Spain |
| Jesús A. González | INAOE, Mexico |
| Manuel Gómez Olmedo | University of Granada, Spain |
| Peter Haddawy | UNU-IIST, Macau |
| Milos Hauskrecht | University of Pittsburg, USA |
| Jesse Hoey | University of Waterloo, Canada |
| Pedro Larrañaga | Technical University of Madrid, Spain |
| Tze-Yun Leong | National University of Singapore, Singapore |
| Subramani Mani | Vanderbilt University, USA |
| Stijn Meganck | Vrije Universiteit Brussel, Belgium |
| Stephen Muggleton | Imperial College London, UK |
| Agnieszka Oniśko | Bialystok Technical University, Poland |
| C. David Page | University of Wisconsin, USA |
| Niels Peek | University of Amsterdam, The Netherlands |
| Mark H. Phillips | University of Washington, USA |

| | |
|-------------------|--|
| Luc De Raedt | KU Leuven, Belgium |
| Alberto Roverato | University of Bologna, Italy |
| Alberto Riva | University of Florida, USA |
| Marco Scutari | University College London, UK |
| Enrique Sucar | INAOE, Mexico |
| Allan Tucker | Brunel University, UK |
| Marina Velikova | Radboud University Nijmegen, The Netherlands |
| Shyam Visweswaran | University of Pittsburg, USA |

Preface

With the current trend toward pervasive health care, personalised health care, and the ever growing amount of evidence coming from biomedical research, methods that can handle reasoning and learning under uncertainty are becoming more and more important. The ongoing developments of the past two decades in the field of artificial intelligence have made it now possible to apply probabilistic methods to solve problems in real-world biomedical domains.

Many representations have been suggested for solving problems in biomedical domains. Bayesian networks and influence diagrams have proved themselves useful for problems where probabilistic uncertainty is important, such as medical decision making and prognostics; logics have proved themselves useful in areas such as diagnosis. In recent years, the field of statistical relational learning has led to new formalisms which integrate probabilistic graphical models and logic. These formalisms provide exciting new opportunities for medical applications as they can be used to learn from structured medical data and reason with them using both logical and probabilistic methods.

Another major theme for this workshop is in the handling of semantic concepts such as space and time in the biomedical domain. Space is an important concept when developing probabilistic models of, e.g., the spread of infectious disease, either in the hospital or in the community at large. Temporal reasoning is especially important in the context of personalised health care. Consider for example the translation of biomedical research that is expected to lead to more complex decision making, e.g., how to optimally select a sequence of drugs targeting biological pathways when treating a malignant tumour. There are strong expectations that such personalised and specific drugs will soon be available in the clinical practice.

We selected eleven papers for full presentation. All these contributions fit the format of the workshop: they develop new approaches for integrating logical and semantical concepts with probabilistic methods or apply existing methods to problems from the biomedical domain. Furthermore, we feel honoured to have Jesse Davis and Milos Hauskrecht as invited speakers. Jesse Davis has made significant contributions in the application of statistical relational learning techniques in the medical domain. Milos Hauskrecht is well-known for his work in the analysis of time-series data (e.g., using POMDPs) in biomedical informatics.

The organisers would like to acknowledge the support from the AIME organisation. We would also like to thank the program committee members for their support and reviewing, which have improved the accepted papers significantly.

Arjen Hommersom and Peter J.F. Lucas
Workshop chairs

Table of Contents

| | |
|--|-----|
| Invited Talk: Conditional outlier detection for clinical alerting..... | 1 |
| <i>Milos Hauskrecht</i> | |
| Gesture Therapy 2.0: Adapting the rehabilitation therapy to the patient progress..... | 3 |
| <i>Héctor Avilés, Roger Luis, Juan Oropeza, Felipe Orihuela-Espina, Ronald Leder, Jorge Hernández-Franco, and Enrique Sucar</i> | |
| On Identifying Significant Edges in Graphical Models..... | 15 |
| <i>Marco Scutari, Radhakrishnan Nagarajan</i> | |
| Learning Multi-Dimensional Bayesian Network Classifiers Using Markov Blankets: A Case Study in the Prediction of HIV Protease Inhibitors | 29 |
| <i>Hanen Borchani, Concha Bielza, and Pedro Larrañaga</i> | |
| Unveiling HIV mutational networks associated to pharmacological selective pressure: a temporal Bayesian approach | 41 |
| <i>Pablo Hernandez-Leal, Alma Rios-Flores, Santiago Ávila-Rios, Gustavo Reyes-Terán, Jesus A. González, Felipe Orihuela-Espina, Eduardo F. Morales, and L. Enrique Sucar</i> | |
| Bayesian data analytic knowledge bases in genetic association studies: Serotonergic and Dopaminergic Polymorphisms in Impulsivity | 55 |
| <i>P. Sarkozy, P. Marx, G. Varga, A. Szekely, Zs. Nemoda, Zs. Demetrovics, M. Sasvari-Szekely, and P. Antal</i> | |
| Invited Talk: Statistical Relational Learning for Clinical Domains..... | 67 |
| <i>Jesse Davis</i> | |
| A Probabilistic Logic of Qualitative Time | 69 |
| <i>Maarten van der Heijden and Peter J.F. Lucas</i> | |
| Effective priors over model structures applied to DNA binding assay data | 83 |
| <i>Nicos Angelopoulos and Lodewyk Wessels</i> | |
| Modelling Inter-practice Variation of Disease Interactions using Multilevel Bayesian Networks..... | 93 |
| <i>Martijn Lappenschaar, Arjen Hommersom, Stefan Visscher, and Peter J.F. Lucas</i> | |
| Towards a Method of Building Causal Bayesian Networks for Prognostic Decision Support | 107 |
| <i>Barbaros Yet, Zane Perkins, William Marsh, and Norman Fenton</i> | |

| | |
|--|-----|
| Cost-effectiveness analysis with influence diagrams | 121 |
| <i>Manuel Arias and Francisco Javier Díez</i> | |
| Impact of Quality of Bayesian Networks Parameters on Accuracy of Medical Diagnostic Systems | 135 |
| <i>Agnieszka Onisko and Marek J. Druzdel</i> | |
| Author Index | 149 |

Conditional outlier detection for clinical alerting

Milos Hauskrecht

Department of Computer Science
University of Pittsburgh

Abstract. In this talk I present a new statistical outlier detection framework for detecting conditional outliers and its application to identification of unusual patient management decisions and clinical alerting. Our hypothesis is that patient-management decisions that are unusual with respect to past patients may be due to errors and that it is worthwhile to raise an alert if such a condition is encountered. The methodology was tested using data obtained from the electronic health records of 4,486 post-cardiac surgical patients and the opinion of a panel of experts. The results support that outlier-based alerting can lead to reasonably low false alert rates and that stronger outliers are correlated with higher alert rates.

Gesture Therapy 2.0: Adapting the rehabilitation therapy to the patient progress

Héctor Avilés¹, Roger Luis¹, Juan Oropeza¹, Felipe Orihuela-Espina¹, Ronald Leder², Jorge Hernández-Franco³, and Enrique Sucar¹

¹ National Institute of Astrophysics, Optics and Electronics (INAOE)
Tonantzintla, Puebla, Mexico

{haviles, rofer_luve, jmanuel, f.oriuela-espina, esucar}@inaoe.mx

² Universidad Nacional Autónoma de México (UNAM)
Ciudad Universitaria, México

{rleder}@ieee.org

³ National Institute of Neurology and Neurosurgery (INNN)
Mexico City, Mexico
{jhfranco}@medicapolanco.com

Abstract. Gesture Therapy is a low-cost, virtual reality based therapy to aid stroke patients to recover upper-limbs' motor skills by playing videogames that re-create daily-life activities. In this paper, we extend our previous work on Gesture Therapy to adjust the difficulty of the game accordingly to the performance of the user. A partially observable Markov decision process is used to adapt the difficulty level from speed and deviation from smooth motion paths. In addition, we consider recently proposed criteria such as *motivation*, *adaptation*, and *task repetition* to design our rehabilitation games. Preliminary results show that the system is able to identify the user dexterity and adjust the difficulty level accordingly. We believe that, in the future, the adaptation module will strongly contribute to relocate therapy sessions from rehabilitation centers to home, and also to decrease the need for on-site assistance of professional therapists.

1 Introduction

In 2010, the American Heart Association –*AHA*– and the National Institutes of Health –*NIH*– branded stroke as the leading cause of motor and cognitive disabilities in the world requiring rehabilitation therapy [1]. Unfortunately, the high cost of therapies [2] makes them prohibitive for many people, especially in middle and low income countries. In a previous work [3] we described the Gesture Therapy system –*GT*– that is a low-cost therapy system that combines computer vision and virtual reality –*VR*– in which patients recover while they play short games that simulate daily-life activities. Initial clinical evaluations [4] suggest that GT can compete with more classical occupational therapies in terms of motor skill recovery and confirms the suspected extra motivation evoked in the patients.

In this paper, we present an extension to the GT platform that incorporates an adaptation module aimed to maintain a suitable level of difficulty of the game accordingly to the user performance. A partially observable Markov decision process –*POMDP*– was designed to infer the patient status from simple motion features such as speed and deviation from smooth motion paths. The action policy generated from the POMDP is able to ease the difficulty level of the game if the performance of the patient decreases, it hardens the game as the user motricity improves, or maintain the difficulty level if no change is observed. In contrast to commercial games in which the difficulty is always increased through the game, GT 2.0 adapts to the patient needs in real-time. Moreover, we argue that our approach fulfill recently proposed criteria to design rehabilitation games such as user motivational feedback, adaptation to motor skill level, task repetition, and simple objective achievement. An initial evaluation shows the reliability of the overall approach, the suitability of the system to identify the user dexterity and to adapt the difficulty of the game accordingly. We believe that, in the future, GT 2.0 will effectively contribute to move therapy sessions from stroke rehabilitation units to home, and to reduce the need for continuous evaluation and assistance of physical therapists.

The outline of this document is as follows. Section 2 discusses related work on recent alternatives to rehabilitation therapy. In Section 3, the general architecture of our system is described. Section 4 presents the POMDP-based adaptation module. Section 5 describes our experiments and results. Finally, section 6 draws our conclusions and future work.

2 Related work

Literature shows several different alternatives to aid patients to improve motor skills and functions of upper-limbs after stroke. Common rehabilitation treatments –or interventions– includes mental practice, neurophysiological approaches, and repetitive task training [5]. Recently, encouraging initial results obtained from computer-based assistive technology such as robotics and VR games have shown the effectiveness and flexibility of these interfaces to aid patients in their rehabilitation process [6–8]. In particular, virtual reality based interventions represent a more affordable option to other therapy alternatives, without compromising the restoration of motor skills. Furthermore, it allows for fast prototyping and development, wide acceptance, and personal customization added to an edge in evoking motivation from the patients [9]. It has been shown that probabilistic-graphical models such as Bayesian networks and POMDPs provide a suitable framework to generate advice to patients and adjust the therapy accordingly to their requirements [10, 11]. In particular, POMDPs account for noisy observations from the world, uncertainty in the current state of the process and its transitions to other states, while computing a mapping from belief states to actions –also called a *policy* [12]. These are desirable features to be considered when observing and modelling the activity of post stroke patients; for example, in the presence of either motion capture sensors with limited

perceptual capabilities or partial descriptions about the current internal state of the patient. These situations may produce non accurate observations of the performance of the user in the game.

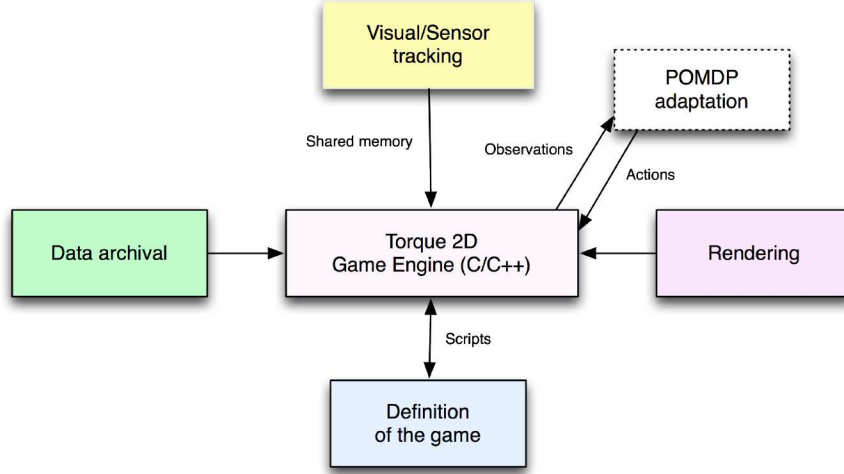


Fig. 1. Schematic representation of the Gesture Therapy 2.0 architecture.

Recently, and closely related to our proposal, Goetschalckx *et al.* [19] described a POMDP-based approach to adjust the difficulty level of a rehabilitation game using a robotic device as the input interface. However, definite probabilistic models for achieving this goal are far from being fully developed, and new model and interface proposals are needed to be explored, for example, to combine personal adaptation and motivation information to produce an integral and comfortable experience for the patient. The present document is a contribution to solve this problem.

3 Architecture of GT 2.0

The GT 2.0 platform is depicted in Figure 1. The central axis of the platform is built around Torque 2D –*T2D*– that it is an commercial engine for rapid game prototyping. T2D centralises communication among all other modules. The module named *Definition of the game* communicates with the T2D by means of control scripts. The *Data archival* module is a database managing system –*DBMS*– for storing patient movements. *Visual/Sensor tracking* module generates visual and gripper data that are accessed by T2D via *shared memory*. The *Rendering* module provides visual feedback to the user accordingly to its performance. As the user interacts with the system, GT 2.0 stores patient movements in the

data archival and transmit these to the game, which responds to the user input and further provides observable data to the adaptive module. Figure 2 shows an example of a real user interacting with the system in a kitchen environment.

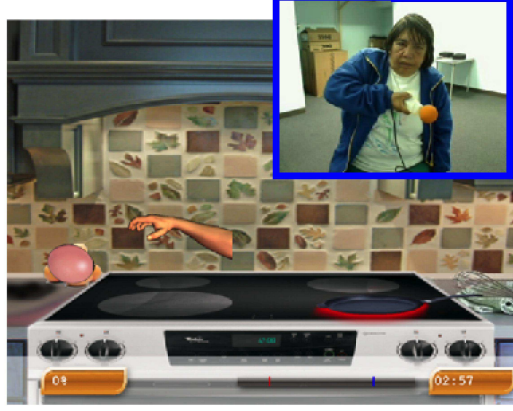


Fig. 2. Example of a real user interacting with the system that re-create a kitchen environment.

Finally, the *POMDP adaptation* module is responsible for changing the difficult level for the game. This module is described in the following section.

4 POMDP adaptation module

As stated above, the GT adaptation module is implemented using a POMDP. The POMDP framework is used to quantify the “convenience” of the states of a system although its real situation is not completely known, and hence, to plan optimal actions to reach a goal state. A POMDP is a tuple

$$POMDP = \langle S, A, O, R, \Omega, T, I \rangle$$

where S is a set of states, A is a set of actions, O is a set of observations, R is a reward function of executing action $a \in A$ when in state $s_t \in S$ and moving to state $s_{t+1} \in S$; T represents the transition probability distribution among states and an action, Ω describes the conditional observation probabilities at each state in S , and finally, I is the initial state probability distribution. Needless to say, a POMDP is one that complies with the Markovian property. A Bayesian graphical representation of a POMDP is illustrated in Figure 3.

Our POMDP implementation is built upon symbolic-Perseus algorithm and software [13] that allow factored representations of state and observation variables. Initially, the POMDP of the GT adaptive module considers two hidden—or state—variables named *Performance* of the user and *Difficulty* of the game;

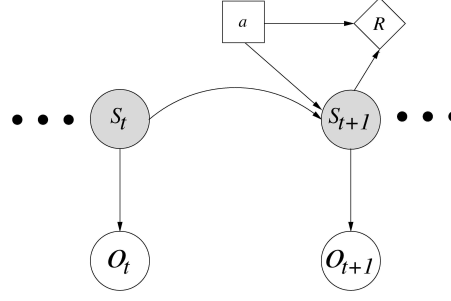


Fig. 3. A Bayesian network description of a POMDP unrolled two times from t to $t+1$, with an action a and a reward function R . Shaded circles indicate state variables S_t and S_{t+1} ; white circles are observation variables O_t and O_{t+1} .

and two observable streams *Control* and *Speed*, as shown in Figure 4. Control is determined as the deviation in the trajectory from a straight movement from origin –cursor position at the instant of target popping– and target location. The more deviation from this straight path, the less control. Control is considered in 3 ranges; low –*coff*–, normal –*cb*– and good –*con*. Speed corresponds to the ratio of distance along the optimum path and execution time. Similarly to control, speed is also considered in 3 ranges; low –*soff*–, normal –*sb*– and good –*son*. The combination of control and speed dictates the performance of the user –*bad*, *good*, and *outstanding*– in turn governing the game difficulty. The variable Difficulty can take three possible values: *easy*, *medium*, and *hard*. We consider three possible actions: *dolvlup*, *dolldown* and *do_nothing* that increases, decreases and keep unchanged the difficulty of the game, respectively. Decisions are made in order to keep the difficulty level in balance with respect to the performance level –i.e., Performance=*bad* and Difficulty=*easy*, Performance=*good* and Difficulty=*medium*, or Performance=*outstanding* and Difficulty=*hard*. For example, if the user performance is outstanding and the game difficulty is easy, the POMDP action policy increases the difficulty of the exercise to medium; if the user continues its excellent performance, the module then increases the game level to hard. Decisions to decrease the difficulty level are made at any moment the performance is below the current difficulty level. Decision to do nothing takes place whenever the performance level is balanced with respect to the difficulty level. Positive rewards are assigned in this latter case only.

5 Experiments and results

In this section, we present a preliminary round of tests performed to evaluate the overall functionality of our proposal and its results. First we discuss design criteria we follow in order to construct our rehabilitation game.

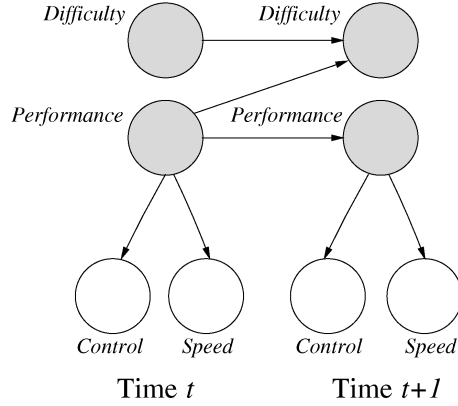


Fig. 4. The GT adaptive module POMDP.

5.1 Design criteria for rehabilitation games

Without detracting of the traditional criteria of repetition and meaning in rehabilitation therapies, the new generation of therapies emphasizes motivation trying to avoid boredom in the therapeutic sessions [8]. Higher motivation by patients often turns into higher commitment thus facilitating recuperation. Serious games used in the motor rehabilitation by means of virtual reality based therapies should provide a safe, adequate and attractive environment to patient [8]. In developing these games not only classical aspects present in leisure games such as temporality, sound, graphics, step or learning curve among other must be cared for. In addition, rehabilitation serious games must also incorporate criteria and objectives specific to rehabilitation of patients with motor impairment. Here, in Table 1 we give a naive classification of a few of these rehabilitation intrinsic criteria. Primary criteria correspond to the elements that ought to be present in any rehabilitation therapy. Secondary criteria agglutinates those criteria desirable to be incorporated in the rehabilitation games. Finally, because the population affected by stroke has a certain profile [1], it is not surprising that these games incorporate elements of entertainment focused to objective population, some of which are covered in Table 1. The table does not aim to be exhaustive and for instance does not include clinical parameters such as timing or dose, the initial state of the patient represented by its age, presence of depression, health and lifestyle or support network, as it does not include particularities of the motivation such as expectancy, self-efficacy, compliance and adherence, etc [5, 8, 14–18]. Yet those criteria indicated in Table 1 can be argued to be the most relevant permeating all other criteria.

To demonstrate the adaptive system a dummy game was created. Note that this game is not part of the GT platform itself as it lack some of the features indicated in Table 1. Nonetheless, the game incorporates the motion replay, and is arguably engaging. The types of movement covered are complex –the movement occurs along the two screen dimensions–, the platform automatically provides

| Primary | Secondary | Social group specific |
|--|---|---|
| <ul style="list-style-type: none"> – Motion replay – Meaningful/Significative task – Motivation | <ul style="list-style-type: none"> – Type of movement covered –unidirectional, combined, complex– – Exhaustion/Fatigue level – Affective level – Reduction of compensation movements – Focus diverted from exercise – Simple interface and clear objective. – Therapy-appropriate range of movement – Adaptability to patient needs | <p>Elderly</p> <ul style="list-style-type: none"> – Control of frustration level – Attention level <p>Infants</p> <ul style="list-style-type: none"> – Educative – Language |

Table 1. Summary of some important criteria for designing serious games focused in motor rehabilitation therapy, with exemplary criteria for social groups. This criteria as well as others not included here have been mentioned earlier in literature [5, 8, 14–18].

support for detecting compensation and the interface is clear and objective. The adaptability criteria is the one under development in this paper. The game presents a flying cooked steak and a cursor. The goal of the game is simply to cross the cursor over the flying steak as many times as possible and as fastly and accurately as possible, using the GT hand grip control. Everytime the heart of the steak is crossed with the cursor the steak randomly changes its location in the screen, with the new location only being allowed to be at a maximum distance from the cursor depending on the game level dictated by the adaptive system.

5.2 Evaluation and results

To test our approach we use a 2Gb 1.86GHz Dual-Core desktop computer. To capture images a cheap webcam was used. Image frame rate of the visual system is about 15 FPS with a resolution of 320×240 pixels. Experiments were performed in our lab environment with artificial light. The distance between the user and the videocamera is around 1m.

We conducted a preliminary set of experiments to evaluate the adaptation module response to modify the difficulty level of the game accordingly to the user performance. Four healthy subjects were recruited among the staff and students of the National Institute of Astrophysics, Optics and Electronics. Following a

brief description of the system by the researcher the subjects were allowed to familiarise with the system for about 1 minute. Each session lasted 3 minutes aprox. corresponding to interleaved blocks of activity –1 minute each– and –rest 20 seconds each–, starting and finishing in activity.



Fig. 5. One of the subjects playing the adaptive game during the experiment.

A depiction of the experimental set up can be seen in Figure 5. As the system is prepared for people with motor disabilities, the healthy subjects can easily master the game and consequently reach the maximum level of the game within a very short amount of time. In this sense, the rest periods are aimed to allow the adaptive system to sense a “bad performance” and thus lower the game difficulty. After the session was finished, the log file was saved for offline analysis. Data from subject 3 was discarded as he deviated from the protocol.

Figure 6 illustrates the timecourse of the the observable variables –speed and control–, the action taken –action– and the output –level– for one of the subjects. From these results it emanates how the POMDP assesses the player ability status and determines in real-time whether to increase, decrease or leave unaffected the game difficulty. During the active periods the game through the POMDP and based on the observation of speed and control takes no action and maintain the level. A few seconds after entering the rest periods, and detecting no success in achieving the task as a result of lack of activity, the POMDP delivers a *down* action which the game translates to a decrease in the level. As the activity resumes, and the player hit the target in appropriate speed and control, the POMDP emit a *up* resulting in the increment in game level. Note how the system determines action to lower the difficulty during the rest periods for all subjects, even if not necessarily resetting to the easiest level. The module reaction in terms of increasing or decreasing game difficulty may be tuned according to the therapy requirements.

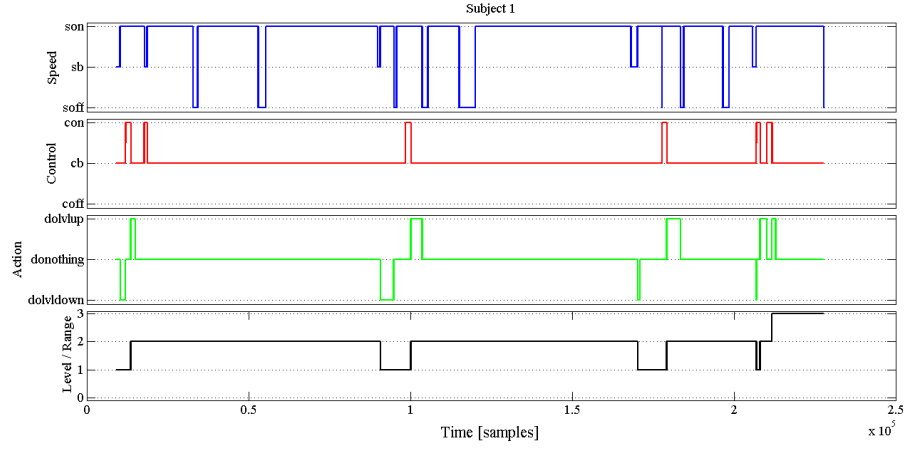


Fig. 6. Timecourse of the observable variables –speed and control–, the hidden variables –performance and difficulty–, the action taken –action– and the output –level– for subject 1. It is shown that the level of the game is modified accordingly to the values of speed and control observations. For example, whenever speed and control variables reflect a good performance –son and con, respectively– the adaptation module selects the action *dolvup*, and the level of the game is increased.

6 Conclusions

In this paper we presented a new adaptation module for the GT rehabilitation platform. The system infers the user performance by observing its speed and control while it plays. Upon determining subject performance, the POMDP decides whether to increase or decrease game difficulty -or leave it unchanged-. We have shown preliminary results for healthy subjects. We intend to further analyse the system response in healthy subjects before carry on with rehabilitation patients. As new games are incorporated to the gesture therapy repository, it will be necessary to assess how the game meets the goals/objectives of a rehabilitation game.

In terms of the rehabilitation therapy, the new module provides a dynamic environment capable of tailoring behaviour to user progress. The importance of this is twofold; (i) the presence of therapist is not required continuously and (ii) the patient can now proceed with its rehabilitation at home at its own pace. This contribution to the rehabilitation therapy is expected to go hand in hand with an enhancement of the user experience. This is yet to be tested in a clinical evaluation but the hypothesis being that the patient will benefit from a self-pace progress resulting in an adequate patient specific cognitive, physiological and mechanical load reduction. Moreover, it is expected that the self-pace progress will increase compliance with the therapy, an especially critical point when therapy is to occur away from the therapist.

References

1. Lloyd-Jones, D., Adams, R.J., Brown, T.M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T.B., Ford, E., Furie, K., Gillespie, C., Go, A., Greenlund, K., Haase, N., Hailpern, S., Ho, M.P. Howard, V. Kissela, B., Kitner, S., Lackland, D., Lisabeth, L., Marelli, A., McDermott, M.M. Meigs, J., Mozaffarian, D., Mussolino, M., Nichol, G., Roger, V.L., Rosamond, W. Sacco, R., Sorlie, P. Stafford, R., Thom, T. Wasserthiel-Smoller, S., Wong, Nathan D., Wylie-Rossett, J., (on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee).: Heart disease and stroke statistics - 2010 Update A report from the American Heart Association. In: *Circulation*, vol. 121, pp. e46-e215. (2010)
2. Lo, A.C., Guarino, P.D., Richards, L.G., Haselkorn, J.K., Wittenberg, G.F., Federman, D.G., Ringer, R.J., Wagner, T.H., Krebs, H.I., Volpe, B.T., Bever, C.T., Bravata, D.M., Duncan, P.W., Corn, B.H., Maffucci, A.D., Nadeau, S.E., Conroy, S.S., Powell, J.M., Huang, G.D., Peduzzi, P.: Robot-assisted therapy for long term upper-limb impairment after stroke. In: *New England Journal of Medicine*, vol. 362/19, pp. 1772–1783. (2010)
3. Sucar, L.E., Luis, R., Leder, R., Hernández, J., Sánchez, I.: Gesture therapy: A vision-based system for upper extremity stroke rehabilitation. In: 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 3690–3693. (2010)
4. Sucar, L. E., Molina, A., Leder, R., and Hernández, J., and Sánchez, I., Gesture therapy: a clinical evaluation. In: 3rd International Conference of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (ICST) on Pervasive Computing Technologies for Healthcare (Pervasive Health 2009), pp. 1–5. (2009)
5. Langhorne, P., Coupar, F., Pollock, A.: Motor recovery after stroke: a systematic review. In: *Lancet Neurology*, vol. 8, pp. 741–754. (2009)
6. Tapus, A., Tapus, C., and Mataric, M.J.: Hands-Off Therapist Robot Behavior Adaptation to User Personality for Post-Stroke Rehabilitation Therapy. In: *International Conference on Robotics and Automation*, pp. 1547–1553. (2007)
7. Kwakkel, G., Boudewijn J. K., Hermano I. K.: Effects of Robot-assisted therapy on upper limb recovery after stroke: A Systematic Review. In: *Neurorehabil Neural Repair*, vol.22/3, pp. 111–121. (2008)
8. Flores, E., Tobon, G., Cavallaro, E., Cavallaro, F.I., Perry, J.C., Keller, T.: Improving Patient Motivation in Game Development for Motor Deficit Rehabilitation. In: *ACE '08 Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, pp. 381–384. (2008)
9. August, K. Bleichenbacher, D. Adamovich, S.: Virtual Reality Physical Therapy: A Telerehabilitation Tool for Hand and Finger Movement Exercise Monitoring and Motor Skills Analysis. In: *Proceedings of the IEEE 31st Annual Northeast Bioengineering Conference*, pp. 73–74. (2005)
10. Suh, M.K., Lee, K.J., Heu, A., Nahapetian, A., Sarrafzadeh M.: Bayesian Networks-Based Interval Training Guidance System for Cancer Rehabilitation. In: *Conference on Mobile Computing, Applications, and Services (MobiCASE)*. (2009)
11. Kan, P., Hoey, J., Mihailidis, A.: Automated upper extremity rehabilitation for stroke patients using a partially observable Markov decision process. In: *Association for Advancement of Artificial Intelligence (AAAI), Fall Symposium on AI in Eldercare*. (2008)

12. Spaan, M.T.J., Vlassis, N.: Perseus: Randomized point-based value iteration for POMDPs. In: *Journal of Artificial Intelligence Research*, vol. 24, pp. 195–220. (2005)
13. Hoey, J., Poupart, P., Bertoldi, A., Craig, T., Boutilier, C., Mihailidis, A.: Automated Handwashing Assistance for Persons with Dementia Using Video and a Partially Observable Markov Decision Process, In: *Computer Vision and Image Understanding (CVIU)*, vol. 114/5, pp. 503–519. (2009)
14. Alklind Taylor, A.-S. Backlund, P. Engstrm, H. Johannesson, M. Lebram, M. Chang, M. (Ed.): Gamers against all odds. In: *Edutainment, Lecture Notes in Computer Science*, vol. 5670, pp. 1–12 Springer-Verlag, (2009)
15. Burke, J. W. McNeill, M. D. J. Charles, D. K. Morrow, P. J. Crosbie, J. H. McDonough, S. M.: Optimising engagement for stroke rehabilitation using serious games. In: *Visual Computer*, England, Aug (2009)
16. Ingles, J. L. Eskes, G. A. Phillips, S. J.: Fatigue after stroke. In: *Archives of Physical Medicine and Rehabilitation*, vol 80, pp. 173–178 (1999)
17. Patton, J. Rymer, Z. (Eds.): Rehabilitation Robotics for Stroke Recovery Machines Assissting Recovery from Stroke - Rehabilitation Engineering Research Center (MARS-RERC) State of the Science (SOS) Meeting. Chicago, USA, May (2011)
18. Prasad, G. Herman, P. Coyle, D. McDonough, S. Crosbie, J.: Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: a feasibility study. In: *Journal of Neuroengineering and Rehabilitation*, vol. 7, pp. 60 (17pp) (2010)
19. Goetschalckx, R., Missura, O., Hoey, J., Gärtner, T.: Games with Dynamic Difficulty Adjustment through the Use of Partially Observable Markov Decision Processes, In: *27th International Conference on Machine Learning, Workshop on Machine Learning and Games*. (2010)

On Identifying Significant Edges in Graphical Models

Marco Scutari¹ and Radhakrishnan Nagarajan²

¹ Genetics Institute, University College London, London, United Kingdom.

² Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA.

Abstract. Graphical models, and in particular Bayesian networks, have been widely used to investigate data in the biological and healthcare domains. This can be attributed to the recent explosion of high-throughput data across these domains and the importance of understanding the causal relationships between the variables of interest. However, classic model validation techniques for identifying significant edges rely on the choice of an ad-hoc threshold, which is non-trivial and can have a pronounced impact on the conclusions of the analysis.

In this paper, we overcome this limitation by proposing simple, statistically-motivated approach based on L_1 approximation for identifying significant edges. The effectiveness of the proposed approach is demonstrated on gene expression data sets across two published experimental studies.

Keywords: graphical models, model averaging, L_1 approximation.

1 Introduction and Background

Graphical models [18, 28] are a class of statistical models which combine the rigour of a probabilistic approach with the intuitive representation of relationships given by graphs. They are composed by a set $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ of *random variables* describing the quantities of interest and a *graph* $\mathcal{G} = (\mathbf{V}, E)$ in which each *vertex* $v \in \mathbf{V}$ is associated with one of the random variables in \mathbf{X} . The *edges* $e \in E$ are used to express the dependence relationships among the variables in \mathbf{X} . The set of these relationships is often referred to as the *dependence structure* of the graph. Different classes of graphs express these relationships with different semantics, which have in common the principle that graphical separation of two vertices implies the conditional independence of the corresponding random variables [28]. The two examples most commonly found in literature are *Markov networks* [8, 35], which use undirected graphs, and *Bayesian networks* [20, 26], which use directed acyclic graphs.

In principle, there are many possible choices for the joint distribution of \mathbf{X} , depending on the nature of the data and the aims of the analysis. However, literature have focused mostly on two cases: the *discrete case* [14, 35], in which both \mathbf{X} and the X_i are multinomial random variables, and the *continuous case* [13, 35], in which \mathbf{X} is multivariate normal and the X_i are univariate normal

random variables. In the former, the parameters of interest are the *conditional probabilities* associated with each variable, usually represented as conditional probability tables; in the latter, the parameters of interest are the *partial correlation coefficients* between each variable and its neighbours in \mathcal{G} .

The estimation of the structure of the graph \mathcal{G} is called *structure learning* [8, 18], and consists in finding the graph structure that encodes the conditional independencies present in the data. Ideally it should coincide with the dependence structure of \mathbf{X} , or it should at least identify a distribution as close as possible to the correct one in the probability space. Several algorithms have been presented in literature for this problem, thanks to the application of many results from probability, information and optimisation theory. Despite differences in theoretical backgrounds and terminology, they can all be traced to only three approaches: *constraint-based* (which are based on conditional independence tests), *score-based* (which are based on goodness-of-fit scores) and *hybrid* (which combine the previous two approaches). For some examples see Bromberg et al. [1], Castelo and Roverato [2], Friedman et al. [12], Larrañaga et al. [21] and Tsamardinos et al. [34].

On the other hand, model validation techniques have not been developed at a similar pace. For example, the characteristics of structure learning algorithms are still studied using a small number of reference data sets [10, 24] as benchmarks, and differences from the true (known) structure are measured with purely descriptive measures such as Hamming distance [17]. This approach is clearly not possible when validating networks learned from real world data sets (because the true structure of their probability distribution is not known) and presents some limits even for synthetic data.

A more systematic approach to model validation, and in particular to the problem of identifying statistically significant features in a network, has been developed by Friedman et al. [11] using bootstrap resampling [9] and model averaging [5]. It can be summarised as follows:

1. For $b = 1, 2, \dots, m$:
 - (a) sample a new data set \mathbf{X}_b^* from the original data \mathbf{X} using either parametric or nonparametric bootstrap;
 - (b) learn the structure of the graphical model $G_b = (\mathbf{V}, E_b)$ from \mathbf{X}_b^* .
2. Estimate the probability that each possible edge e_i , $i = 1, \dots, k$ is present in the true network structure $\mathcal{G}_0 = (\mathbf{V}, E_0)$ as

$$\hat{P}(e_i) = \frac{1}{m} \sum_{b=1}^m \mathbb{1}_{\{e_i \in E_b\}}, \quad (1)$$

where $\mathbb{1}_{\{e_i \in E_b\}}$ is the indicator function of the event $\{e_i \in E_b\}$ (i.e., it is equal to 1 if $e_i \in E_b$ and 0 otherwise).

The empirical probabilities $\hat{P}(e_i)$ are known as *edge intensities* or *arc strengths*, and can be interpreted as the degree of *confidence* that e_i is present in the

network structure \mathcal{G}_0 describing the true dependence structure of \mathbf{X} ³. However, they are difficult to evaluate, because the probability distribution of the networks \mathcal{G}_b in the space of the network structures is unknown. As a result, the value of the confidence threshold (i.e. the minimum degree of confidence for an edge to be significant and therefore accepted as an edge of \mathcal{G}_0) is an unknown function of both the data and the structure learning algorithm. This has proved to be a serious limitation in the identification of significant edges and has led to the use of ad-hoc, pre-defined thresholds in spite of the impact on model validation evidenced by several studies [11, 15]. An exception is Nagarajan et al. [25], whose approach will be discussed below.

Apart from this limitation, Friedman’s approach is very general and can be used in a wide range of settings. First of all, it can be applied to any kind of graphical model with only minor adjustments (for example, accounting for the direction of the edges in Bayesian networks). Furthermore, it does not require any distributional assumption on the data in addition to the ones needed to by the structure learning algorithm. No assumption is made on the latter, either, so any score-based, constraint-based or hybrid algorithm can be used.

In this paper, we propose a statistically-motivated estimator for the confidence threshold minimising the L_1 norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the confidence levels of the unknown network \mathcal{G}_0 . Subsequently, we demonstrate the effectiveness of the proposed approach by re-investigating two experimental data sets from Nagarajan et al. [25] and Sachs et al. [30].

2 Selecting Significant Edges

Consider the empirical probabilities $\hat{P}(e_i)$ defined in Eq. 1, and denote them with $\hat{\mathbf{p}} = \{\hat{p}_i, i = 1, \dots, k\}$. For a graph of size n , $k = n(n - 1)/2$. Furthermore, consider the order statistic

$$\hat{\mathbf{p}}_{(\cdot)} = \{0 \leq \hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \dots \leq \hat{p}_{(k)} \leq 1\} \quad (2)$$

derived from $\hat{\mathbf{p}}$. It is intuitively clear that the first elements of $\hat{\mathbf{p}}_{(\cdot)}$ are more likely to be associated with non-significant edges, and that the last elements of $\hat{\mathbf{p}}_{(\cdot)}$ are more likely to be associated with significant edges. The ideal configuration $\tilde{\mathbf{p}}_{(\cdot)}$ of $\hat{\mathbf{p}}_{(\cdot)}$ would be

$$\tilde{p}_{(i)} = \begin{cases} 1 & \text{if } e_{(i)} \in E_0 \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

that is the set of probabilities that characterises any edge as either significant or non-significant without any uncertainty. In other words,

$$\tilde{\mathbf{p}}_{(\cdot)} = \{0, \dots, 0, 1, \dots, 1\}. \quad (4)$$

³ The probabilities $\hat{P}(e_i)$ are in fact an estimator of the expected value of the $\{0, 1\}$ random vector describing the presence of each possible edge in \mathcal{G}_0 . As such, they do not sum to one and are dependent on one another in a nontrivial way.

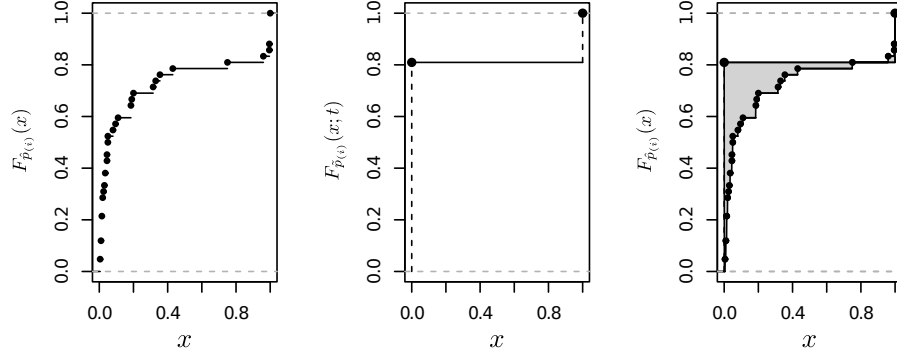


Fig. 1. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$ (left), the cumulative distribution function $F_{\tilde{\mathbf{p}}_{(\cdot)}}(x; t)$ (centre) and the L_1 norm between the two (right).

Such a configuration arises from the limit case in which all the networks \mathcal{G}_b have exactly the same structure. This may happen in practise with a consistent structure learning algorithm when the sample size is large [4, 22].

A useful characterisation of $\hat{\mathbf{p}}_{(\cdot)}$ and $\tilde{\mathbf{p}}_{(\cdot)}$ can be obtained through the empirical cumulative distribution functions of the respective elements,

$$F_{\hat{\mathbf{p}}_{(\cdot)}}(x) = \frac{1}{k} \sum_{i=1}^k \mathbb{1}_{\{\hat{p}_{(i)} < x\}} \quad (5)$$

and

$$F_{\tilde{\mathbf{p}}_{(\cdot)}}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0) \\ t & \text{if } x \in [0, 1) \\ 1 & \text{if } x \in [1, +\infty) \end{cases}. \quad (6)$$

In particular, t corresponds to the fraction of elements of $\tilde{\mathbf{p}}_{(\cdot)}$ equal to zero and is a measure of the fraction of non-significant edges. At the same time, t provides a threshold for separating the elements of $\tilde{\mathbf{p}}_{(\cdot)}$, namely

$$e_{(i)} \in E_0 \iff \hat{p}_{(i)} > F_{\tilde{\mathbf{p}}_{(\cdot)}}^{-1}(t). \quad (7)$$

More importantly, estimating t from data provides a statistically motivated threshold for separating significant edges from non-significant ones. In practise, this amounts to approximating the ideal, asymptotic empirical cumulative distribution function $F_{\tilde{\mathbf{p}}_{(\cdot)}}$ with its finite sample estimate $F_{\hat{\mathbf{p}}_{(\cdot)}}$. Such an approximation can be computed in many different ways, depending on the norm used to measure the distance between $F_{\hat{\mathbf{p}}_{(\cdot)}}$ and $F_{\tilde{\mathbf{p}}_{(\cdot)}}$ as a function of t . Common choices are the L_p family of norms [19], which includes the Euclidean norm, and Csiszar's f -divergences [7], which include Kullback-Leibler divergence.

The L_1 norm

$$L_1(t; \hat{\mathbf{p}}_{(\cdot)}) = \int |F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\hat{\mathbf{p}}_{(\cdot)}}(x; t)| dx \quad (8)$$

appears to be particularly suited to this problem; an example is shown in Fig. 1. First of all, note that $F_{\hat{\mathbf{p}}_{(\cdot)}}$ is piecewise constant, changing value only at the points $\hat{p}_{(i)}$; this descends from the definition of empirical cumulative distribution function. Therefore, for the problem at hand Eq. 8 simplifies to

$$L_1(t; \hat{\mathbf{p}}_{(\cdot)}) = \sum_{x_i \in \{\{0\} \cup \hat{\mathbf{p}}_{(\cdot)} \cup \{1\}\}} |F_{\hat{\mathbf{p}}_{(\cdot)}}(x_i) - t| (x_{i+1} - x_i), \quad (9)$$

which can be computed in linear time from $\hat{\mathbf{p}}_{(\cdot)}$. Its minimisation is also straightforward using linear programming [27]. Furthermore, compared to the more common L_2 norm

$$L_2(t; \hat{\mathbf{p}}_{(\cdot)}) = \int [F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\hat{\mathbf{p}}_{(\cdot)}}(x; t)]^2 dx \quad (10)$$

or the L_∞ norm

$$L_\infty(t; \hat{\mathbf{p}}_{(\cdot)}) = \max_{x \in [0,1]} \{|F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\hat{\mathbf{p}}_{(\cdot)}}(x; t)|\}, \quad (11)$$

the L_1 norm does not place as much weight on large deviations, making it robust against a wide variety of configurations of $\hat{\mathbf{p}}_{(\cdot)}$.

Then the identification of significant edges can be thought of either as a *least absolute deviations estimation* or an L_1 *approximation* of the form

$$\hat{t} = \operatorname{argmin}_{t \in [0,1]} L_1(t; \hat{\mathbf{p}}_{(\cdot)}) \quad (12)$$

followed by the application of the following rule:

$$e_{(i)} \in E_0 \iff \hat{p}_{(i)} > F_{\hat{\mathbf{p}}_{(\cdot)}}^{-1}(\hat{t}). \quad (13)$$

A simple example of its use is illustrated below.

Example 1. Consider a graphical model based on an undirected graph \mathcal{G} with vertex set $\mathbf{V} = \{A, B, C, D\}$. The set of possible edges of \mathcal{G} contains 6 elements: (A, B) , (A, C) , (A, D) , (B, C) , (B, D) and (C, D) . Suppose that that we have estimated the following confidence values:

$$\hat{p}_{AB} = 0.2242, \quad \hat{p}_{AC} = 0.0460, \quad \hat{p}_{AD} = 0.8935, \quad (14)$$

$$\hat{p}_{BC} = 0.3921, \quad \hat{p}_{BD} = 0.7689, \quad \hat{p}_{CD} = 0.9439. \quad (15)$$

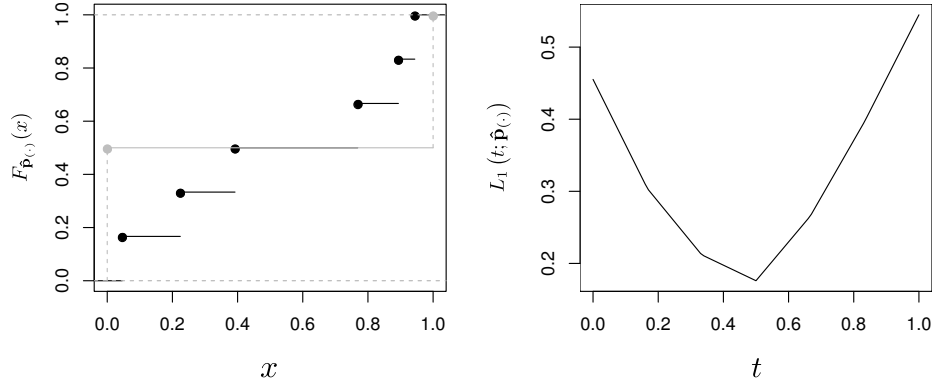


Fig. 2. The cumulative distribution functions $F_{\hat{\mathbf{p}}_{(\cdot)}}$ and $F_{\hat{\mathbf{p}}_{(\cdot)}}(\hat{t})$, respectively in black and grey (left), and the $L_1(t; \hat{\mathbf{p}}_{(\cdot)})$ norm (right) from Example 1.

Then $\hat{\mathbf{p}}_{(\cdot)} = \{0.0460, 0.2242, 0.3921, 0.7689, 0.8935, 0.9439\}$ and

$$F_{\hat{\mathbf{p}}_{(\cdot)}}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0.0460) \\ \frac{1}{6} & \text{if } x \in [0.0460, 0.2242) \\ \frac{2}{6} & \text{if } x \in [0.2242, 0.3921) \\ \frac{3}{6} & \text{if } x \in [0.3921, 0.7689) \\ \frac{4}{6} & \text{if } x \in [0.7689, 0.8935) \\ \frac{5}{6} & \text{if } x \in [0.8935, 0.9439) \\ 1 & \text{if } x \in [0.9439, +\infty) \end{cases} \quad (16)$$

The L_1 norm takes the form

$$\begin{aligned} L_1(t; \hat{\mathbf{p}}_{(\cdot)}) = & |0 - t|(0.0460 - 0) + \left| \frac{1}{6} - t \right| (0.2242 - 0.0460) + \\ & \left| \frac{2}{6} - t \right| (0.3921 - 0.2242) + \left| \frac{3}{6} - t \right| (0.7689 - 0.3921) + \\ & \left| \frac{4}{6} - t \right| (0.8935 - 0.7689) + \left| \frac{5}{6} - t \right| (0.9439 - 0.8935) + \\ & |1 - t|(1 - 0.9439) \end{aligned} \quad (17)$$

and is minimised for $\hat{t} = 0.4999816$. Therefore, an edge is deemed significant if its confidence is strictly greater than $F_{\hat{\mathbf{p}}_{(\cdot)}}^{-1}(0.4999816) = 0.3921$, or, equivalently, if it has confidence of at least 0.7689; only (A, D) , (B, D) and (C, D) satisfy this condition.

3 Applications to Gene Expression Profiles

We will now examine the effectiveness of the proposed estimator for the significance threshold on two gene expression data sets from Nagarajan et al. [25] and Sachs et al. [30]. All the analyses will be performed with the bnlearn package [31, 32] for R [29], which implements several methods for structure learning, parameter estimation and inference on Bayesian networks. Following Imoto et al. [16], we will consider the edges of the Bayesian networks disregarding their direction. Edges identified as significant will be oriented according to the direction observed with the highest frequency in the bootstrapped networks \mathcal{G}_b . This combined approach allows the proposed estimator to handle the edges whose direction cannot be determined by the structure learning algorithm (which are called *score equivalent edges* [3]), because directions are completely ignored in the estimation. At the same time, it can be observed that in practise the two possible orientations of such edges usually appear with comparable frequencies in the networks \mathcal{G}_b . Therefore, proper interpretation of their meaning in the network structure resulting from the application of the approach outlined in Sec. 2 is possible.

3.1 Differentiation Potential of Aged Myogenic Progenitors

In a recent study [25] the interplay between crucial myogenic (Myogenin, Myf-5, Myo-D1), adipogenic (C/EBP α , DDIT3, FoxC2, PPAR γ), and Wnt-related genes (Lrp5, Wnt5a) orchestrating aged myogenic progenitor differentiation was investigated by Nagarajan et al. using clonal gene expression profiles in conjunction with Bayesian network structure learning techniques. The objective was to investigate possible functional relationships between these diverse differentiation programs reflected by the edges in the resulting networks. The clonal expression profiles were generated from RNA isolated across 34 clones of myogenic progenitors obtained across 24-month-old mice and real-time RT-PCR was used to quantify the gene expression. Such an approach implicitly accommodates inherent uncertainty in gene expression profiles and justified the choice of probabilistic models.

In the same study, the authors proposed a non-parametric resampling approach to identify significant functional relationships. Starting from Friedman’s definition of confidence levels (Eq. 1), they computed the *noise floor distribution* $\hat{\mathbf{f}} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_k\}$ of the edges by randomly permuting the expression of each gene and performing Bayesian network structure learning on the resulting data sets. An edge e_i was deemed significant if $\hat{p}_i > \max(\hat{\mathbf{f}})$. In addition to revealing several functional relationships documented in literature, the study also revealed new relationships that were immune to the choice of the structure learning techniques. These results were established across clonal expression data normalised using three different housekeeping genes and networks learned with three different structure learning algorithms.

The approach presented in [25] has two important limitations. First, the computational cost of generating the noise floor distribution may discourage its

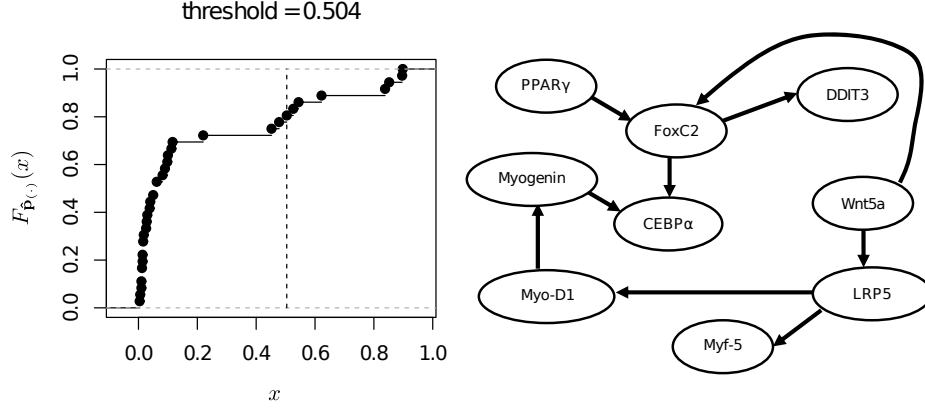


Fig. 3. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$ for the myogenic progenitors data from Nagarajan et al. [25] (on the left), and the network structure resulting from the selection of the significant edges (on the right). The vertical dashed line in the plot of $F_{\hat{\mathbf{p}}_{(\cdot)}}$ represents the threshold $F_{\hat{\mathbf{p}}_{(\cdot)}}^{-1}(\hat{t})$.

application to large data sets. In fact, the generation of the required permutations of the data and the subsequent structure learning (in addition to the bootstrap resampling and the subsequent learning required for the estimation of $\hat{\mathbf{p}}$) essentially doubles the computational complexity of Friedman’s approach. Second, a large sample size may result in an extremely low value of $\max(\hat{\mathbf{f}})$, and therefore in a large number of false positives.

In the present study, we re-investigate the myogenic progenitor clonal expression data normalised using housekeeping gene GAPDH with the approach outlined in Sec. 2 and a constraint-based learning strategy based on the Incremental Association Markov Blanket (IAMB) algorithm [33]. The latter is used to learn the Markov blanket of each vertex as a preliminary step to reduce the number of its candidate parents and children; a network structure satisfying these constraints is then identified as in the Grow-Shrink algorithm [23]. It is important to note that this strategy was also used in the original study [25], hence its choice. The order statistic $\hat{\mathbf{p}}_{(\cdot)}$ was computed from 500 bootstrap samples. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$, the estimated threshold and the network with the significant edges are shown in Fig. 3.

All edges identified as significant in the earlier study [25] across the various structure learning techniques and normalisations techniques were also identified by the proposed approach (see Fig. 3D in [25]). In contrast to Fig. 3, the original study using IAMB and normalisations with respect to GAPDH alone detected a considerable number of additional edges (see Fig. 3A in [25]). Thus it is quite possible that the approach proposed in this paper reduces the number of false positives and spurious functional relationships between the genes. Furthermore, the application of the proposed approach in conjunction with the algorithm from

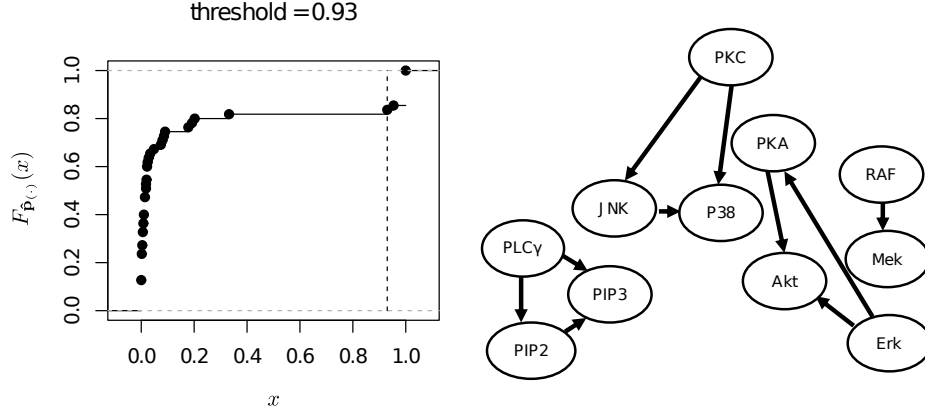


Fig. 4. The empirical cumulative distribution function of $\hat{\mathbf{p}}(\cdot)$ for the flow cytometry data from Sachs et al. [30] (on the left), and the network structure resulting from the selection of the significant edges (on the right). The vertical dashed line in the plot of $F_{\hat{\mathbf{p}}(\cdot)}$ represents the threshold $F_{\hat{\mathbf{p}}(\cdot)}^{-1}(\hat{t})$.

Imoto et al. [16] reveals directionality of the edges, in contrast to the undirected network reported by Nagarajan et al. [25].

3.2 Protein Signalling in Flow Cytometry Data

In a recent study, Sachs et al. [30] used Bayesian networks as a tool for identifying causal influences in cellular signalling networks from simultaneous measurement of multiple phosphorylated proteins and phospholipids across single cells. The authors used a battery of perturbations in addition to the unperturbed data to arrive at the final network representation. A greedy search score-based algorithm that maximises the posterior probability of the network [14] and accommodates for variations in the joint probability distribution across the unperturbed and perturbed data sets was used to identify the edges [6]. More importantly, significant edges were selected using an arbitrary significance threshold of 0.85 (see Fig. 3, [30]). A detailed comparison between the learned network and functional relationships documented in literature was presented in the same study.

We investigate the performance of the proposed approach in identifying significant functional relationships from the same experimental data. However, we limit ourselves to the data recorded without applying any molecular intervention, which amount to 854 observations for 11 variables. We compare and contrast our results to those obtained using an arbitrary threshold of 0.85. The combination of perturbed and non-perturbed observations studied in Sachs et al. [30] cannot be analysed with our approach, because each subset of the data follows a different probability distribution and therefore there is no single “true” network \mathcal{G}_0 . Analysis of the unperturbed data using the approach presented in Sec. 2 reveals the edges reported in the original study. The resulting network is shown in Fig.

4 along with $F_{\hat{\mathbf{p}}_{(i)}}$ and the estimated threshold. From the plot of $F_{\hat{\mathbf{p}}_{(i)}}$ we can clearly see that significant and non-significant edges present widely different levels of confidence, to the point that any threshold between 0.4 and 0.9 results in the same network structure. This, along with the value of the estimated threshold ($\hat{p}_{(i)} \geq 0.93$), shows that the noisiness of the data relative to the sample size is low. In other words, the sample is big enough for the structure learning algorithm to reliably select the significant edges. The edges identified by the proposed method were the same as those identified by [30] using general stimulatory cues excluding the data with interventions (see Fig. 4A in [30], Supplementary Information). In contrast to [30], using Imoto et al. [16] approach in conjunction with the proposed thresholding method we were able to identify the direction of the edges in the network. The directionality correlated with functional relationships documented in literature (Tab. 3, [30], Supplementary Information) as well as with the directionality of the network learned from both perturbed and unperturbed data (Fig. 3, [30]).

4 Conclusions

Network abstractions provided by graphical models have enjoyed considerable attention across the biological and medical communities, where they are used to represent the concerted working as a system as opposed to independent entities. For example, these networks may represent the underlying signalling mechanisms and pathways within the context of biological data. Classic model validation techniques identify significant edges using an ad-hoc threshold across multiple realisations of networks learned from the given data. Such ad-hoc approaches can have pronounced effect on the resulting networks and biological conclusions. The present study overcomes this critical caveat by proposing a more straightforward and statistically-motivated approach for identifying significant edges in a graphical model. The proposed estimator minimises the L_1 norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the “edge confidence” determined from the given data. The effectiveness of the proposed approach is demonstrated on gene expression data sets across two different studies [25, 30]. However, the approach is defined in a more general setting and can be applied to many classes of graphical models learned from any kind of data. A more detailed investigation is underway in elucidating the various aspects of the proposed approach.

Acknowledgements

This work was supported by the National Library of Medicine grant R03LM008853 (Radhakrishnan Nagarajan) and BBSRC & Technology Board grant TS/I002170/1 (Marco Scutari). Marco Scutari would also like to thank Adriana Brogini for proofreading the paper and providing useful suggestions.

Bibliography

- [1] Bromberg, F., Margaritis, D., Honavar, V.: Efficient Markov Network Structure Discovery using Independence Tests. *Journal of Artificial Intelligence Research* 35, 449–485 (2009)
- [2] Castelo, R., Roverato, A.: A Robust Procedure For Gaussian Graphical Model Search From Microarray Data With p Larger Than n . *Journal of Machine Learning Research* 7, 2621–2650 (2006)
- [3] Chickering, D.M.: A Transformational Characterization of Equivalent Bayesian Network Structures. In: *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI95)*. pp. 87–98 (1995)
- [4] Chickering, D.M.: Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research* 3, 507–554 (2002)
- [5] Claeskens, G., Hjort, N.L.: *Model Selection and Model Averaging*. Cambridge University Press (2008)
- [6] Cooper, G.F., Yoo, C.: Causal Discovery from a Mixture of Experimental and Observational Data. In: *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI)*. pp. 116–125. Morgan Kaufmann (1999)
- [7] Csiszár, I., Shields, P.: *Information Theory and Statistics: A Tutorial*. Now Publishers Inc. (2004)
- [8] Edwards, D.I.: *Introduction to Graphical Modelling*. Springer, 2nd edn. (2000)
- [9] Efron, B., Tibshirani, R.: *An Introduction to the Bootstrap*. Chapman & Hall (1993)
- [10] Elidan, G.: *Bayesian Network Repository* (2001), <http://www.cs.huji.ac.il/site/labs/compbio/Repository>
- [11] Friedman, N., Goldszmidt, M., Wyner, A.: Data Analysis with Bayesian Networks: A Bootstrap Approach. In: *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*. pp. 206 – 215. Morgan Kaufmann (1999)
- [12] Friedman, N., Pe’er, D., Nachman, I.: Learning Bayesian Network Structure from Massive Datasets: The “Sparse Candidate” Algorithm. In: *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI)*. pp. 206–221. Morgan Kaufmann (1999)
- [13] Geiger, D., Heckerman, D.: *Learning Gaussian Networks*. Tech. rep., Microsoft Research, Redmond, Washington (1994), Available as Technical Report MSR-TR-94-10
- [14] Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20(3), 197–243 (September 1995), Available as Technical Report MSR-TR-94-09
- [15] Husmeier, D.: Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks, vol. 19 (2003)

- [16] Imoto, S., Kim, S.Y., Shimodaira, H., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S.: Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression. *Genome Informatics* 13, 369–370 (2002)
- [17] Jungnickel, D.: *Graphs, Networks and Algorithms*. Springer-Verlag, 3rd edn. (2008)
- [18] Koller, D., Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press (2009)
- [19] Kolmogorov, A.N., Fomin, S.V.: *Elements of the Theory of Functions and Functional Analysis*. Graylock Press (1957)
- [20] Korb, K., Nicholson, A.: *Bayesian Artificial Intelligence*. Chapman and Hall (2004)
- [21] Larrañaga, P., Sierra, B., Gallego, M.J., Michelena, M.J., Picaza, J.M.: Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma. In: *Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe (AIME '97)*. pp. 261–272. Springer (1997)
- [22] Lauritzen, S.L.: *Graphical Models*. Oxford University Press (1996)
- [23] Margaritis, D.: *Learning Bayesian Network Model Structure from Data*. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA (May 2003), Available as Technical Report CMU-CS-03-153
- [24] Murphy, P., Aha, D.: *UCI Machine Learning Repository* (1995), <http://archive.ics.uci.edu/ml>
- [25] Nagarajan, R., Datta, S., Scutari, M., Beggs, M.L., Nolen, G.T., Peterson, C.A.: Functional Relationships Between Genes Associated with Differentiation Potential of Aged Myogenic Progenitors. *Frontiers in Physiology* 1(21), 1–8 (2010)
- [26] Neapolitan, R.E.: *Learning Bayesian Networks*. Prentice Hall (2003)
- [27] Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer-Verlag (1999)
- [28] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann (1988)
- [29] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2010), <http://www.R-project.org>
- [30] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science* 308(5721), 523–529 (2005)
- [31] Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35(3), 1–22 (2010)
- [32] Scutari, M.: *bnlearn: Bayesian Network Structure Learning* (2011), <http://www.bnlearn.com/>, R package version 2.4
- [33] Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Algorithms for Large Scale Markov Blanket Discovery. In: *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference*. pp. 376–381. AAAI Press (2003)

- [34] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning* 65(1), 31–78 (2006)
- [35] Whittaker, J.: *Graphical Models in Applied Multivariate Statistics*. Wiley (1990)

Learning Multi-Dimensional Bayesian Network Classifiers Using Markov Blankets: A Case Study in the Prediction of HIV Protease Inhibitors

Hanen Borchani, Concha Bielza, and Pedro Larrañaga

Computational Intelligence Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte, 28660, Spain.
hanen.borchani@upm.es, mcbielza@fi.upm.es, pedro.larranaga@fi.upm.es

Abstract. Multi-dimensional Bayesian network classifiers (MBCs) are Bayesian network classifiers especially designed to solve multi-dimensional classification problems, where each instance in the data set has to be assigned to one or more class variables. In this paper, we introduce a new method for learning MBCs from data basically based on determining the Markov blanket around each class variable using the HITON algorithm. Our method is applied to the human immunodeficiency virus (HIV) protease inhibitor prediction problem. The experimental study showed promising results in terms of classification accuracy, and we gained insight from the learned MBC structure into the different possible interactions among protease inhibitors and resistance mutations.

1 Introduction

Multi-dimensional classification is an extension of the classical one-dimensional classification, where each instance given by a vector of m features $\mathbf{x} = (x_1, \dots, x_m)$ is associated with a vector of d class values $\mathbf{c} = (c_1, \dots, c_d)$ rather than a single class value [16]. Recently, the concept of multi-dimensionality has been introduced in Bayesian network classifiers providing an accurate modelling of this emerging problem and ensuring interactions among all variables [4, 5, 9, 16–18]. In these probabilistic graphical models, known as multi-dimensional Bayesian network classifiers (MBCs), the graphical structure partitions the set of class and feature variables into three different subgraphs: class subgraph, feature subgraph and bridge subgraph, and the parameter set defines the conditional probability distribution of each variable given its parents.

In this paper, we introduce a novel MBC learning algorithm based on Markov blankets. Motivated by the fact that the classification is unaffected by parts of the structure that lie outside the Markov blankets of the class variables, we first build the Markov blanket around each class variable using the well-known HITON algorithm [1–3], and then we determine edge directionality over all three MBC subgraphs. Thanks to this filter and local approach to MBC learning, we can lighten the computational burden of MBC learning using wrapper algorithms [4, 5, 16] and provide more accurate MBC structures.

We finally apply our Markov blanket MBC (MB-MBC) algorithm to the problem of predicting human immunodeficiency virus (HIV) protease inhibitors (PIs) given an input set of resistance mutations that an HIV patient carries. In general, a combination of several antiretroviral PI drugs should be repeatedly administered for each patient in order to prevent and treat the HIV infection. We analyze a data set obtained from the Stanford HIV protease database [13]. The class variables are eight protease inhibitor drugs (i.e., $d=8$) and the feature variables are 74 predefined mutations [10] associated with resistance to protease inhibitors (i.e., $m=74$). Experimental results were promising in terms of classification accuracy as well as of the identification of interactions among drugs and resistance mutations, which were either consistent with the current knowledge or not previously mentioned in the literature.

The remainder of this paper is organized as follows. Section 2 introduces Bayesian networks. Section 3 presents MBCs and briefly reviews state-of-the-art MBC learning algorithms. Section 4 describes our new MBC learning approach. Section 5 presents the experimental study on the HIV protease inhibitor data set. Finally, Section 6 sums up the paper with some conclusions.

2 Background

A Bayesian network over a set of discrete random variables $\mathbf{U} = \{X_1, \dots, X_n\}$, $n \geq 1$, is a pair $\mathcal{B} = (\mathcal{G}, \Theta)$. $\mathcal{G} = (V, A)$ is a directed acyclic graph (DAG) whose vertices V correspond to variables in \mathbf{U} and whose arcs A represent direct dependencies between the vertices. Θ is a set of conditional probability distributions such that $\theta_{x_i | \mathbf{pa}(x_i)} = p(x_i | \mathbf{pa}(x_i))$ defines the conditional probability of each possible value x_i of X_i given a set value $\mathbf{pa}(x_i)$ of $\mathbf{Pa}(X_i)$, where $\mathbf{Pa}(X_i)$ denotes the set of parents of X_i in \mathcal{G} .

A Bayesian network \mathcal{B} represents a joint probability distribution over \mathbf{U} factorized according to structure \mathcal{G} as follows:

$$p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | \mathbf{Pa}(X_i)). \quad (1)$$

Definition 1. *Conditional Independence.* Two variables X and Y are conditionally independent given \mathbf{Z} , denoted as $I(X, Y | \mathbf{Z})$, iff $P(X | Y, \mathbf{Z}) = P(X | \mathbf{Z})$ for all values x, y, \mathbf{z} of X, Y, \mathbf{Z} , respectively, such that $P(\mathbf{Z} = \mathbf{z}) > 0$.

Definition 2. *A Markov blanket of a variable X , denoted as $MB(X)$, is a minimal set of variables with the following property: $I(X, \mathbf{S} | MB(X))$ holds for every variable subset \mathbf{S} with no variables in $MB(X) \cup X$.*

In other words, $MB(X)$ is a minimal set of variables conditioned by which X is conditionally independent of all the remaining variables. Under the faithfulness assumption, $MB(X)$ consists of the union of the set of parents, children, and parents of children (i.e., spouses) of X [11].

3 Multi-dimensional Bayesian Network Classifiers

In this section we present MBCs, then briefly review the state-of-the-art methods for learning these models from data.

Definition 3. A multi-dimensional Bayesian network classifier is a Bayesian network $\mathcal{B} = (\mathcal{G}, \Theta)$ where the structure $\mathcal{G} = (V, A)$ has a restricted topology. The set of n vertices V is partitioned into two sets: $V_C = \{C_1, \dots, C_d\}, d \geq 1$, of class variables and $V_X = \{X_1, \dots, X_m\}, m \geq 1$, of feature variables ($d + m = n$). The set of arcs A is partitioned into three sets A_C , A_X and A_{CX} , such that:

- $A_C \subseteq V_C \times V_C$ is composed of the arcs between the class variables having a subgraph $\mathcal{G}_C = (V_C, A_C)$ -class subgraph- of \mathcal{G} induced by V_C .
- $A_X \subseteq V_X \times V_X$ is composed of the arcs between the feature variables having a subgraph $\mathcal{G}_X = (V_X, A_X)$ -feature subgraph- of \mathcal{G} induced by V_X .
- $A_{CX} \subseteq V_C \times V_X$ is composed of the arcs from the class variables to the feature variables having a subgraph $\mathcal{G}_{CX} = (V, A_{CX})$ -bridge subgraph- of \mathcal{G} connecting class and feature variables.

Depending on the graphical structures of the class and feature subgraphs MBCs can be divided into several families. These families can be denoted as **class subgraph structure-feature subgraph structure** MBCs, where the possible structures of each subgraph are: empty, tree, polytree, or DAG [4].

Classification with an MBC under a 0-1 loss function is equivalent to solving the most probable explanation (MPE) problem, i.e., for a given fact $\mathbf{x} = (x_1, \dots, x_m)$ we have to obtain

$$\begin{aligned} \mathbf{c}^* &= (c_1^*, \dots, c_d^*) \\ &= \arg \max_{c_1, \dots, c_d} p(C_1 = c_1, \dots, C_d = c_d \mid \mathbf{x}). \end{aligned} \quad (2)$$

Example 1. An example of an MBC structure is shown in Figure 1. V_C contains four classes, V_X includes seven features, and the structure \mathcal{G} is equal to $\mathcal{G}_C \cup \mathcal{G}_X \cup \mathcal{G}_{CX}$. We have

$$\begin{aligned} \max_{c_1, \dots, c_d} p(C_1 = c_1, \dots, C_d = c_d \mid \mathbf{x}) &\propto \max_{c_1, \dots, c_d} p(c_1 \mid c_2, c_3) p(c_2) p(c_3) p(c_4) \\ &\quad \cdot p(x_1 \mid c_2, x_4) p(x_2 \mid c_1, c_2) p(x_3 \mid c_4) p(x_4 \mid c_1) \\ &\quad \cdot p(x_5 \mid c_4) p(x_6 \mid c_3, x_3, x_7) p(x_7 \mid c_4, x_3). \end{aligned}$$

Several approaches have recently been proposed to learn MBCs from data. In [16], Van der Gaag and de Waal use Chow and Liu's algorithm [6] to learn the class and feature subgraphs of a **tree-tree** MBC, then they greedily select the bridge subgraph, using a wrapper method, aiming to induce the most accurate classifier. De Waal and Van der Gaag later presented a theoretical approach for learning **polytree-polytree** MBCs in [9]. Class and feature subgraphs are separately generated using Rebane and Pearl's algorithm [12]; however, the induction of the bridge subgraph was not specified.

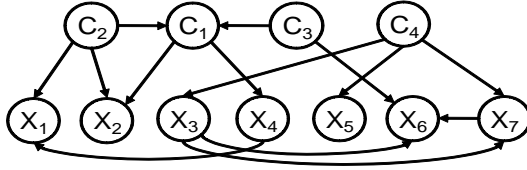


Fig. 1. An example of an MBC structure.

More recently, a two-step method was proposed by Zaragoza et al. [17] to also learn **polytree-polytree** MBCs. First, they build class and feature subgraphs using Chow and Liu’s algorithm [6] and generate an initial bridge subgraph based on mutual information. Then, in a second step, they refine the bridge subgraph by adding more arcs to improve MBC accuracy.

Bielza et al. [4] propose three MBC learning algorithms: pure filter (guided by any filter algorithm based on a fixed ordering among the variables), pure wrapper (guided by the classification accuracy) and a hybrid algorithm (a combination of pure filter and pure wrapper). Note that none of these algorithms places any constraints on the subgraph structures of the generated MBCs.

In [5], we propose a learning algorithm for class-bridge decomposable MBCs, instead of general MBCs, based on a greedy forward selection wrapper approach. Class or feature subgraphs can have any type of structure. Compared with prior algorithms in [4, 9, 16], our method performs better and requires less computational time than the existing wrapper algorithms.

Moreover, Zaragoza et al. present a two-step method in [18]. In the first phase, a tree-based Bayesian network that represents the dependency relations between the class variables is learned. In the second phase, several chain classifiers are built using selective naive Bayes models, such that the order of the class variables in the chain is consistent with the class subgraph. At the end, the results of the different generated orders are combined in a final ensemble model.

4 Learning Multi-Dimensional Bayesian Network Classifiers Using Markov Blankets

In this section we describe a new algorithm for learning MBCs from data based on Markov blanket discovery. Our objective is to tackle the shortcomings of our previous learning method [5], mainly its high computational cost, by taking advantage of the merits of a filter approach. This should considerably lighten the computational burden, especially when the data set includes a large number of class and feature variables, while guaranteeing good performance.

Additionally, this work is motivated by its application to the HIV drug resistance problem, where it is not only important to build an MBC with a high predictive power but also to discover the resistance pathways of each HIV drug by analyzing the MBC structure. Applying our previous learning method [5] may not always lead to an accurate MBC structure, since arcs between features

are selected at random in the feature subgraph learning steps. This may affect the overall quality of the learned MBC structure and lead consequently to misinterpretations.

To deal with this issue, we make use of Markov blankets. In recent years, several specialized Markov blanket learning methods have been proposed in the literature, such as GS, TPDA, IAMB and its variants, MMHC, MMMB and HITON (see [2, 3] and their references for reviews). In this paper, we only consider and apply the HITON algorithm [1–3] in the context of multi-dimensional Bayesian network classifiers. In fact, the HITON algorithm was empirically proven to outperform most of the state-of-the-art Markov blanket discovery algorithms in terms of combined classification performance and feature set parsimony [2].

The idea of our Markov blanket MBC (MB-MBC) learning algorithm is simple and consists of applying the HITON algorithm to each class variable and then specifying directionality over the MBC subgraphs. HITON identifies the Markov blanket of each class variable in a two-phase scheme, HITON-MB and HITON-PC, outlined respectively in Algorithms 1 and 2.

Step 1 of HITON-MB identifies the parents and children of each class variable C_i , denoted $PC(C_i)$, by calling the HITON-PC algorithm. Then, it determines the PC set for every member T of $PC(C_i)$ (steps 2 to 4). The Markov blanket set $MB(C_i)$ is initialized with $PC(C_i)$ (step 5) and set \mathbf{S} includes potential spouses of C_i (step 6). From steps 7 to 14, HITON-MB loops over all members of \mathbf{S} to identify correct spouses of C_i . $MB(C_i)$ is finally returned in step 15.

Algorithm 1 HITON-MB(C_i)

```

1.  $PC(C_i) \leftarrow \text{HITON-PC}(C_i)$ 
2. for every variable  $T \in PC(C_i)$  do
3.    $PC(T) \leftarrow \text{HITON-PC}(T)$ 
4. end for
5.  $MB(C_i) \leftarrow PC(C_i)$ 
6.  $\mathbf{S} \leftarrow \{\bigcup_{T \in PC(C_i)} PC(T)\} \setminus \{PC(C_i) \cup C_i\}$ 
7. for every variable  $X \in \mathbf{S}$  do
8.   Retrieve a subset  $\mathbf{Z}$  s.t.  $I(X, C_i \mid \mathbf{Z})$ 
9.   for every variable  $T \in PC(C_i)$  s.t.  $X \in PC(T)$  do
10.    if  $\neg I(X, C_i \mid \mathbf{Z} \cup \{T\})$  then
11.      Insert  $X$  into  $MB(C_i)$ 
12.    end if
13.  end for
14. end for
15. return  $MB(C_i)$ 
```

HITON-PC starts with an empty set of candidates $PC(T)$, ranks the variables X in OPEN by priority of inclusion according to $I(X, T)$ and discards variables having $I(X, T) = 0$. Then, for every new variable inserted into $PC(T)$, it checks if there is any variable inside $PC(T)$ that is independent of T given some subset \mathbf{Z} . In this case, this variable will be removed from $PC(T)$ (steps 6 to 11). These steps are iterated until there are no more variables in OPEN. Finally, $PC(T)$ is filtered using the symmetry criterion (steps 13 to 17). In fact, for every $X \in$

$PC(T)$, the symmetrical relation holds iff $T \in PC(X)$. Otherwise, i.e., if $T \notin PC(X)$, X will be removed from $PC(T)$. At the end of this step, we obtain $PC(T)$ [2].

Algorithm 2 HITON-PC(T)

```

1.  $PC(T) \leftarrow \emptyset$ 
2.  $OPEN \leftarrow \mathbf{U} \setminus \{T \cup PC(T)\}$ 
3. Sort the variables  $X$  in  $OPEN$  in descending order according to  $I(X, T)$ 
4. Remove from  $OPEN$  variables  $X$  having  $I(X, T) = 0$ 
5. repeat
6.   Insert at end of  $PC(T)$  the first variable in  $OPEN$  and remove it from  $OPEN$ 
7.   for every variable  $X \in PC(T)$  do
8.     if  $\exists \mathbf{Z} \subseteq PC(T) \setminus \{X\}$ , s.t.  $I(X, T | \mathbf{Z})$  then
9.       Remove  $X$  from  $PC(T)$ .
10.    end if
11.  end for
12. until  $OPEN = \emptyset$ 
13. for every variable  $X \in PC(T)$  do
14.   if  $T \notin PC(X)$  then
15.     Remove  $X$  from  $PC(T)$ 
16.   end if
17. end for
18. return  $PC(T)$ .
```

Note that the complexity of both algorithms could be controlled using a parameter max_{CS} restricting the maximum number of elements in the conditioning sets \mathbf{Z} [2]. In our experiments, we use the G^2 statistical test to evaluate the conditional independencies between variables with a threshold significance level of $\alpha = 0.05$, and we consider different values of $max_{CS} = 1, 2, 3, 4, 5$.

Unlike the HITON algorithm that only determines the Markov blanket of a single target variable for solving the variable selection problem, our algorithm considers many target variables, then induces the MBC graphical structure. Given the MBC definition, direct parents of any class variable C_i , $i = 1, \dots, d$, can only be among the remaining class variables, whereas direct children or spouses of C_i can include either class or feature variables. We can then easily deduce the different MBC subgraphs based on the results of the HITON algorithm:

- *Class subgraph*: we firstly insert an edge between each class variable C_i and any class variable belonging to its corresponding parents-children set $PC(C_i)$. Then, we direct all these edges using the PC algorithm [15].
- *Bridge subgraph*: this is built by inserting an arc from each class variable C_i to every feature variable belonging to $PC(C_i)$.
- *Feature subgraph*: for every feature X in the set $MB(C_i) \setminus PC(C_i)$, i.e., for every spouse X , we insert an arc from X to the corresponding common child given by $PC(X) \cap PC(C_i)$. Moreover, more arcs can be added especially to discover additional dependency relationships among features. In fact, for every feature X , child of C_i , we determine the set $\mathbf{Y} = PC(X) \setminus (\{C_i\} \cup \{MB(C_i) \cap PC(X)\})$. If $\mathbf{Y} \neq \emptyset$, we insert an arc from X to every feature variable in \mathbf{Y} .

5 Experimental Study

5.1 Data set

Treatments for human immunodeficiency virus (HIV) mostly involve 18 antiretroviral drugs grouped into three classes: nucleoside and nucleotide reverse transcriptase inhibitors (NRTIs) including seven drugs, non-nucleoside reverse transcriptase inhibitors (NNRTIs) including three drugs, and protease inhibitors (PIs) containing eight drugs. In this paper, we studied PIs only, but we plan to extend our study to both NRTIs and NNRTIs in the future.

We analyzed a data set obtained from the Stanford HIV protease database [13] containing antiretroviral PI treatment histories from 1255 patients. These treatment histories were collected from previously published studies. Eight PI drugs (i.e., $d=8$) are considered: Atazanavir (ATV), Darunavir (DRV), Fosamprenavir (FPV), Indinavir (IDV), Lopinavir (LPV), Nelfinavir (NFV), Saquinavir (SQV) and Tipranavir (TPV). There may be one or multiple isolates for the same patient. Each isolate corresponds to a sample in the data set, including a list of resistance mutations and a combination of PIs administered to a patient at a specified time point during his or her course of PI treatment. Only samples where no drug was administered were discarded. Accordingly, the final data set contained a total of 4341 samples. However, the number of PI combinations is not evenly represented; in fact, there are 3256 samples including only 1 PI, 862 samples including 2 PIs, 213 samples including 3 PIs and only 10 samples containing 4 PIs.

Moreover, we considered established drug resistance mutations that were defined in the last International AIDS Society-USA resistance mutation list [10]. The total number of mutations in the protease gene associated with resistance to PIs is 74 (i.e., $m=74$), where 23 are classified as major and the remaining as minor mutations. *Major mutations* are defined as mutations selected first in the presence of the drug or mutations substantially reducing drug susceptibility. *Minor mutations* generally emerge later than major mutations and by themselves do not have a substantial effect [10].

PI drug combinations (respectively resistance mutations) were represented using binary vectors such that every value indicates either the presence, 1, or absence, 0, of an individual PI drug (respectively an individual resistance mutation) in the corresponding sample of the data set. Using a multi-dimensional Bayesian network classifier learned from this data we were able to predict antiretroviral combination PI therapies given sets of input mutations. Thanks to its graphical structure, we were also able to investigate dependencies among classes (i.e., PI drugs), features (i.e., mutations) and between classes and features (i.e., interactions between PI drugs and mutations).

5.2 Experimental Results

We compare our MB-MBC algorithm with what is defined as a multiple classifier method, where each classifier is learned independently (sometimes called binary relevance in the literature on multi-label classification) using the same HITON

approach with just a single class variable. In order to evaluate the performance of the learned MBCs, five 10-fold cross-validation experiments are run for each classifier and each conditioning set size value, i.e., with $max_{CS} = 1, 2, 3, 4, 5$. We use two performance metrics [4], namely:

- The *mean accuracy* over the d class variables:

$$Acc_m = \frac{1}{d} \sum_{i=1}^d \frac{1}{N} \sum_{l=1}^N \delta(c'_{li}, c_{li}), \quad (3)$$

where N is the size of the test set, c'_{li} is the C_i class value predicted by the MBC for sample l , and c_{li} denotes its corresponding real value. $\delta(c'_{li}, c_{li}) = 1$ if the predicted and real class values are equal, i.e., $c'_{li} = c_{li}$, and 0 otherwise.

- The *global accuracy* over the d -dimensional class variable:

$$Acc_g = \frac{1}{N} \sum_{l=1}^N \delta(\mathbf{c}'_l, \mathbf{c}_l). \quad (4)$$

In this case, the vector of predicted classes \mathbf{c}'_l is compared to the vector of real classes \mathbf{c}_l , so that we have $\delta(\mathbf{c}'_l, \mathbf{c}_l) = 1$ if there is a complete equality between both vectors, i.e., $\mathbf{c}'_l = \mathbf{c}_l$, and 0 otherwise.

Table 1 shows the prediction results with mean values and standard deviations for each metric and each method. Note that the best results are obtained with $max_{CS} = 1$ (94% mean accuracy and 71% global accuracy), and as max_{CS} grows, the overall mean and global accuracies decrease. As expected, without exception, MB-MBC outperforms the independent classifier model notably with respect to global accuracy.

Table 1. Estimated performance metrics (mean \pm standard deviation).

| max_{CS} | MB-MBC | | Independent classifiers | |
|------------|---------------------|---------------------|-------------------------|---------------------|
| | Mean accuracy | Global accuracy | Mean accuracy | Global accuracy |
| 1 | 0.9416 \pm 0.0049 | 0.7188 \pm 0.0250 | 0.9339 \pm 0.0019 | 0.7035 \pm 0.0054 |
| 2 | 0.9330 \pm 0.0033 | 0.6868 \pm 0.0075 | 0.9247 \pm 0.0017 | 0.5994 \pm 0.0185 |
| 3 | 0.9193 \pm 0.0031 | 0.6338 \pm 0.0083 | 0.9156 \pm 0.0039 | 0.4960 \pm 0.0153 |
| 4 | 0.8890 \pm 0.0108 | 0.5153 \pm 0.0321 | 0.8775 \pm 0.0091 | 0.4071 \pm 0.0296 |
| 5 | 0.8641 \pm 0.0201 | 0.4266 \pm 0.0568 | 0.8438 \pm 0.0107 | 0.3551 \pm 0.0328 |

In addition, we examined the graphical structure of the most accurate learned MBC, shown in Figure 2, in order to evaluate the usefulness of the proposed learning algorithm in identifying the different interactions between drugs and mutations in the HIV protease data set.

Firstly, the learned network, specifically the class subgraph (red arcs), shows dependency relationships between the following drugs IDV, ATV, NFV, LPV and SQV, which may reveal the extent of cross-resistance between each related pair of these drugs. Notice that, for IDV, which has associations with LPV, ATV

and NFV, Rhee et al. [14] recently proved in their PIs cross-resistance study that IDV and LPV are among the most strongly correlated PIs. In fact, these two drugs had a correlation coefficient value equal to 0.57 [14]. Similarly, based on their study, IDV and ATV, ATV and NFV as well as NFV and IDV had high correlation coefficients. Nevertheless, correlation coefficients between LPV and both drugs NFV and SQV were lower, equal to 0.14 and 0.05 respectively. This goes to confirm then that the dependency relationships identified in the network among the above PI drugs are consistent with Rhee et al.’s study [14].

However, our results were less conclusive for other drugs (DRV, FPV and TPV) since no associations are detected between them or between them and the other drugs. A possible explanation is the lack of available data, as there were fewer than 30 samples for each of these drugs. On this ground, we would require a larger and diverse data set for our future analysis in order to investigate possible interactions between these drugs and the other variables in the network.

Concerning relationships between PI drugs and mutations, visualized by the bridge subgraph (blue arcs), let us first discuss the two possible types of mutations, major and minor, and then how their associations with PI drugs have been previously interpreted in the literature in the context of Bayesian networks. As Defroche et al. found [7, 8], a major mutation actually plays a key role in drug resistance, and thus, should have an unconditional dependency on the drug, and this is indicated in the network graphical structure by the presence of an arc between the major mutation and the drug.

In contrast, a minor mutation further increases drug resistance mostly only in the presence of major mutations. Thus, it is expected to be conditionally independent of the drug but dependent on other major resistance mutations. This is indicated in the network by the presence of an arc between major or minor mutations instead of an arc between the minor mutation and the drug node. Even so, as claimed by Defroche et al. [7], a minor mutation may still be connected to the drug.

Notice that the conditional independencies revealed in our bridge subgraph in Figure 2 are largely consistent with the above definitions, since most of the major mutations are directly connected to one or more drug nodes. For instance, on the left, D30N (which is defined in [10] as a major mutation of NFV) was not only associated with NFV but also with IDV, LPV and SQV, proving again the extent of cross-resistance between these drugs. Similarly, on the right, L76V (which is defined in [10] as a major mutation of LPV) was directly associated with LPV, SQV and NFV. At the center bottom of the network, G48V (major mutation of SQV [10]) was directly associated with SQV and NFV. L90M (another major mutation of SQV [10]) was also directly associated with SQV. I47A, I50L, V82A, V82L, defined in [10] as major mutations of LPV, ATV, IDV and TPV, respectively, were directly associated with the right drugs in the MBC graphical structure.

An important number of minor mutations were also directly connected to drug nodes. L10I and L33F seem to be the main minor mutations: they have the highest number of connections (3) with PI drugs, followed by the minor

mutations L10F and I54V. L10I was associated with IDV, NFV and SQV; L33F with LPV, IDV and NFV; L10F with ATV and IDV, and I54V with LPV and NFV. Additionally, consistently with the latest knowledge in [10], more minor mutations, namely V82A/T, I84V, N88D/S, were associated directly with NFV. Also in agreement with [10], the minor mutation K20R was associated with LPV and the minor mutation I84V was associated with SQV.

From the feature subgraph (green arcs) of the learned MBC we were able to identify interactions among different protease mutations. The mutations with the greatest number of dependency relationships were L10I (21 connections: L10F, L10R, K20R, D30N, M46L, M46I, K43T, G48V, I50V, F53L, I54A, I54T, I62V, A71I, A71V, G73S, V82A, I84V, I85V, L90M, I93L), L10F (15 connections: L10I, L10V, V11I, K20T, L33F, M46I, G48V, I54L, I54V, L63P, I84V, I85V, N88D, L89V, L90M), M46I (8 connections: L10F, L10I, K20I, V32I, M46L, I64L, V77I, N88S), and 7 connections for L33F (L10F, K43T, M46L, I50V, I54L, A71L, V82L) and G48V (L10F, L10I, L24I, D30N, I54A, I54S, V77I).

Finally, of the 19 mutations that present no interactions with other drugs or features (at the bottom), only three are major ones, namely T74P, V82F and N83D. As they have no dependency relationships with any drug, these mutations are completely irrelevant.

6 Conclusion

This paper proposed a novel MBC learning approach using Markov blankets, then presented its application to the HIV protease inhibitors prediction problem. A preliminary experimental analysis showed that our approach performed well and confirmed current knowledge about different interactions among PI drugs and their resistance mutations.

In the near future, we intend to carry out a more extensive experimental study including the comparison of our approach with state-of-the-art MBC learning algorithms, using additional synthetic and real data sets in order to prove the merits of our approach. As regards the HIV drug prediction problem, we plan to apply our approach to the other two HIV drug groups: NRTI and NNRTI. Similarly, two MBCs could be learned separately for each group. However, it would be more interesting to build a single MBC including all the drugs in the PI, NRTI and NNRTI categories. This way, we will be able not only to investigate interactions among drugs and resistance mutations belonging to the same group but also to identify the potential inter-group interactions.

Acknowledgements

The authors would like to thank Dr. Carlos Toro Rueda researcher at Hospital Carlos III de Madrid for his valuable comments. This work has been supported by projects TIN2010-20900-C04-04, Consolider Ingenio 2010-CSD2007-00018, Cajal Blue Brain, and Dynamo (FONCICYT, European Union and Mexico). Hanen Borchani is supported by an FPI fellowship from the Spanish Ministry of Science and Innovation (BES-2008-003901).

References

1. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: HITON: A novel Markov blanket algorithm for optimal variable selection. *AMIA*, 21-25 (2003)
2. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research* 11, 171-234 (2010)
3. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part II: Analysis and extensions. *Journal of Machine Learning Research* 11, 235-284 (2010)
4. Bielza, C., Li, G., Larrañaga, P.: Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, In press, doi: 10.1016/j.ijar.2011.01.007 (2011)
5. Borchani, H., Bielza, C., Larrañaga, P.: Learning CB-decomposable multi-dimensional Bayesian network classifiers. *In Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, 25-32 (2010)
6. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462-467 (1968)
7. Deforche, K., Silander, T., Camacho, R., Grossman, Z., Soares, M.A. et al.: Analysis of HIV-1 pol sequences using Bayesian networks: Implications for drug resistance. *Bioinformatics* 22(24), 2975-2979 (2006)
8. Deforche, K., Camacho, R., Grossman, Z., Silander, T., Soares, M.A. et al.: Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors. *Infection, Genetics and Evolution* 7, 382-390 (2007)
9. De Waal, P.R., van der Gaag, L.C.: Inference and learning in multi-dimensional Bayesian network classifiers. *ECSQARU*, 4724:501-511 (2007)
10. Johnson, V. A., Brun-Vezinet, F., Clotet, B., Gunthard, H.F., Kuritzkes, D.R., et al.: Update of the drug resistance mutations in HIV-1: December 2010. *International AIDS Society-USA, Topics in HIV Medicine* 18(5), 156-163 (2010)
11. Pearl, J., Verma, T.S.: Equivalence and synthesis of causal models. *In Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, 220-227 (1990)
12. Rebane, G., Pearl, J.: The recovery of causal polytrees from statistical data. *UAI*, 222-228 (1989)
13. Rhee, S.Y., Gonzales, M.J., Kantor, R., Betts, J., Ravela, J., Shafer, R.W.: Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* 31(1), 298-303 (2003)
14. Rhee, S.Y., Taylor, J., Fessel, W.J., Kaufman, D., Towner, W. et al.: HIV-1 protease mutations and protease inhibitor cross-resistance. *Antimicrobial Agents and Chemotherapy* 54(10), 4253-4261 (2010)
15. Spirtes, P., Glymour, C. and Scheines, R.: *Causation, Prediction, and Search*. MIT Press, 2nd edition, Cambridge, MA (2000)
16. van der Gaag, L.C., de Waal, P.R.: Multi-dimensional Bayesian network classifiers. *In Proceedings of the Third European Conference on Probabilistic Graphical Models*, 107-114 (2006)
17. Zaragoza, J.H., Sucar, L.E., Morales, E.F.: A two-step method to learn multidimensional Bayesian network classifiers based on mutual information measures. *In Proceedings of the Twenty-Fourth International FLAIRS Conference*, 644-649 (2011)
18. Zaragoza, J.H., Sucar, L.E., Morales, E.F., Larrañaga, P., Bielza, C.: Bayesian chain classifiers for multidimensional classification. *IJCAI*, In press (2011)

Unveiling HIV mutational networks associated to pharmacological selective pressure: a temporal Bayesian approach

Pablo Hernandez-Leal¹, Alma Rios-Flores¹, Santiago Ávila-Rios², Gustavo Reyes-Terán², Jesus A. González¹, Felipe Orihuela-Espina¹, Eduardo F. Morales¹, and L. Enrique Sucar¹

¹ National Institute of Astrophysics, Optics and Electronics (INAOE)
Tonantzintla, Puebla, Mexico

{pablohl,alma.rios,jagonzalez,f.orihuela-espina,emorales,esucar}@inaoe.mx

² Infectious Diseases Research Center at the National Institute of Respiratory Diseases (CIENI-INER)
Mexico City, Mexico
{santiago.avila}@cieni.org.mx

Abstract. Much of the HIV (Human Immunodeficiency Virus) success is due to its evolving capabilities. Understanding viral evolution and its relation to pharmacology is of utmost importance in fighting diseases caused by the HIV. Although the mutations conferring drug resistance are mostly known, the dynamics of the appearance chain of those mutations remains poorly understood. Here we apply a Temporal Nodes Bayesian Network (TNBN) to data extracted from the HIV Stanford database to explore the probabilistic relationships between mutations and antiretrovirals. We aim to unveil existing mutation networks and establish their probabilistic formation sequence. The model predictive accuracy is evaluated in terms of relative Brier score, relative time error and total number of intervals. Robustness of the model is hinted by consistency between two model instances. The learned models capture known relationships, qualitatively providing some measure of validity. Finally, two previously unseen mutational networks are retrieved and their probabilistic temporal sequentiation uncovered. We demonstrate the usefulness of TNBN for studying drug-mutation and mutation-mutation networks and expect to impact the combat against HIV infection by facilitating better treatment planning.

1 Introduction

Viral evolution is an important aspect of the epidemiology of viral diseases such as influenza, hepatitis and human immunodeficiency virus (HIV). This evolution greatly impacts the development of successful vaccines and antiviral drugs, as mutations bestowing drug resistance or immune escape often develop early after the virus is placed under selective pressure. In HIV, this is particularly relevant as the virus ranks among the fastest evolving organisms [7]. Its remarkable viral

replication capability is coupled with a high mutation rate and a high probability of recombination in the viral genome during its replication cycle. These features allow HIV to boast a wide genetic variability even considering only the viral population within a given host. The elevated variation capability of HIV gives the virus a remarkable ability to adapt to multiple selective pressures, including the immune response and antiretroviral therapy. This intra-host genetic variation raises several questions about viral evolution, for example: How much of this diversity is shaped by the selection of the immune response and how much by the antiretroviral therapy? What is the relationship between genetic diversity and clinical outcome? And finally, is it feasible to steer the evolution of HIV in order to reduce drug resistance? Motivated by this last question, it would be desirable to develop proactive therapies that predict the advent of mutations, ergo reducing the risk of drug resistance, rather than waiting for the virus to develop resistance to reactively change the antiretroviral regimen. If we could predict the most likely evolution of the virus in any host, then it would be plausible to select an appropriate antiretroviral regimen that prevents the appearance of mutations, effectively increasing HIV control.

In this work, a Temporal Node Bayesian Networks (TNBN) model was developed to assess the occurrence of probabilistic associations among protease mutations and protease inhibitor drugs. Results of the learning of the model are presented. Our goal was to explain mutational networks in HIV evolution in the viral protease. Our probabilistic graphical model was able to predict antiretroviral drug-associated mutational pathways in the protease gene, revealing the co-occurrence of mutations and its temporal relationships. The technical challenge was to develop a model expressive enough to capture the biological complexity, yet simple enough to allow for a quick interpretation of results. The use of TNBNs for exposing mutational networks is, as far as the authors are aware of, an additional novelty to this work.

The rest of the paper is organized as follows. Section 2 highlights some important notions regarding HIV and how it develops drug resistance. Section 3 justifies the use of TNBN over other existing graphical probabilistic approaches. Section 4 describes the TNBN model. Section 5 presents the experiments and results obtained. Finally, section 6 summarizes the findings and indicates future lines of research.

2 HIV and its defense against antiretroviral therapy

HIV is the causing agent of the disease known as Acquired Immunodeficiency Syndrome (AIDS), a condition in which progressive failure of the immune system allows opportunistic life-threatening infections to occur. HIV is a virus with relatively recent introduction to human populations [17] representing a huge global burden to human health (UNAIDS HIV Global Report 2010). Its structure is schematically depicted in Figure 1.

The replication cycle of HIV is characterized by a reverse-transcription step of the viral RNA genome to a double-stranded DNA molecule, which is then

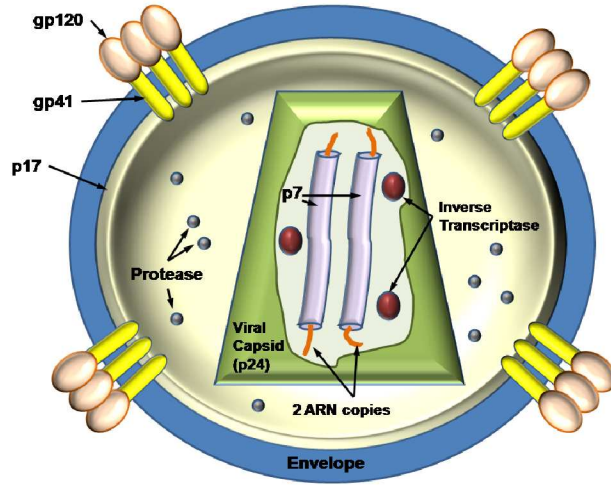


Fig. 1. Schematic representation of the HIV structure. The protease enzyme helping in maturation is illustrated. This enzyme is the target of the protease inhibitors antiretrovirals which are studied in this paper. Some important proteins, e.g. gp120, gp41, p17 and p7 have also been illustrated.

inserted into the host cell genome. To combat HIV infection several antiretroviral (ARV) drugs belonging to different drug classes that affect specific steps in the viral replication cycle have been developed. Antiretroviral therapy (ART) generally consists of well-defined combinations of three or four ARV drugs. Due to its remarkable variation capabilities, HIV can rapidly adapt to the selective pressure imposed by ART through the development of drug resistance mutations, that are fixed in the viral population within the host in known mutational pathways. The development of drug resistant viruses compromises HIV control, with a consequent further deterioration of the patient's immune system. Many of these ARV drug resistance mutations reduce HIV susceptibility to ARV drugs by themselves, while others need to accumulate in order to cause resistance. Moreover, the appearance of some drug resistance mutations implies high costs for viral replication capacity. These costs in viral replication capacity are often compensated by the appearance of additional mutations known as compensating mutations. Due to their polymorphic nature the frequency of compensating mutations can vary between viruses circulating in different geographic areas, making it relevant to study HIV mutational networks in the context of different infected populations.

Abundant literature exists describing computational models aimed to better understand HIV evolution and immunopathogenesis. A good portion of this data is devoted to predict phenotypic HIV resistance to antiretroviral drugs using different approaches such as decision trees [2] or neural networks [6]. Other works try to identify relevant associations between clinical variables and HIV

disease [18]. Surprisingly, among this wealth of literature, works aimed towards the identification of temporal relationships among mutations and drugs in HIV is almost lacking.

In [3] association rules between clinical variables and the failure of the treatment are extracted, they used 15 clinical variables from 8000 patients from data collected since 1981. The results obtained are temporal rules that have as antecedent the increasing of a subset of clinical variables and as consequent the failure of the treatment, given by side effects of the drugs or by the elevated viral count (unsuccessful therapy). None of clinical variables considered are VIH mutations.

3 Bayesian Networks

Information in clinical databases is more often than not imprecise, incomplete, and with errors (noisy), and Bayesian Networks (BNs) [16] are particularly well suited to deal with uncertainty. BNs study probabilistic dependencies and relationships among domain entities. BNs models admit visual representation as a graph consisting of nodes and edges facilitating their analysis and interpretation. Nodes represent random variables and edges represent probabilistic dependencies. This graphical representation is easily understood by humans. An additional advantage is the availability of several methods to learn BN from data, e.g. [14].

BNs have proven to be successful in various domains such as medicine [15] and bioinformatics [20]. However, classical BNs are not well equipped to deal with temporal information. Dynamic Bayesian Networks (DBNs) evolved to tackle this shortcoming [5]. DBNs can be seen as multiple slices of a *static* BN over time, with temporal relations captured as links between adjacent slices. In a DBN, a base model is cloned for each time stage. These copies are linked via the so-called transition network. In this transition network is common that only links between consecutive stages are allowed. Whenever variable changes occur infrequently, the explicit representation of DBNs becomes unnecessarily overexpressive. The alternative are TNBNs [1].

4 Temporal Nodes Bayesian Networks

In a TNBN, each node, known as *temporal node* (TN), represents a random variable that may be in a given state i.e. value interval, throughout the different temporal intervals associated to it. An arc between two temporal nodes describes a temporal probabilistic relation. In TNBNs, each variable (node) represents an event or state change. So, only one (or a few) instance(s) of each variable is required, assuming there is one (or a few) change(s) of a variable state in the temporal range of interest. No copies of the model are needed, thus compacting the representation without losing expressiveness.

The TNBN [1, 9] is composed by a set of TNs connected by arcs representing a probabilistic relationship between TNs. A TN, v_i , is a random variable characterized by a set of states \mathbf{S} . Each state is defined by an ordered pair

$S = (\lambda, \tau)$, where λ is the particular value taken by v_i during its associated interval $\tau = [a, b]$, corresponding to the time interval in which the state change, i.e. change in value, occurs. In addition, each TN contains an extra default state $s = (\text{'no change'}, \emptyset)$ with no associated interval. Time is discretized in a finite number of intervals, allowing a different number and duration of intervals for each node (multiple granularity). Each interval defined for a child node represents the possible delays between the occurrence of one of its parent events and the corresponding child event. If a node lacks defined intervals for all its states then it is referred to as *instantaneous node*. There is at most one state change for each variable (TN) in the temporal range of interest.

Formally, let \mathbf{V} be a set of temporal and instantaneous nodes and \mathbf{E} a set of arcs between nodes, a TNBN is defined as:

Definition 1. A TNBN is a pair $B = (G, \Theta)$ where G is a directed acyclic graph, $G = (\mathbf{V}, \mathbf{E})$ and, Θ is a set of parameters quantifying the network. Θ contains the values $\Theta_{v_i} = P(v_i | Pa(v_i))$ for each $v_i \in \mathbf{V}$; where $Pa(v_i)$ represents the set of parents of v_i in G .

The learning algorithm for TNBN used in this work has been presented in [11]. Briefly, the learning algorithm is described:

1. First, it performs an initial discretization of the temporal variables, for example using an Equal-Width discretization. With this process it obtains an initial approximation of the intervals for all the Temporal Nodes.
2. Then it performs a standard BN structural learning, the algorithm uses the K2 learning algorithm [4], to obtain an initial structure. This structure will be used in the third step, the interval learning algorithm.
3. The interval learning algorithm refines the intervals for each TN by means of clustering. For this, it uses the information of the configurations of the parent nodes. To obtain some intervals a clustering algorithm for the temporal data is used. The approach uses a Gaussian mixture model. Each cluster corresponds, in principle, to a temporal interval. The intervals are defined in terms of the μ and the σ of the clusters. The algorithm obtains different sets of intervals that are merged and combined, these process will generate different interval sets that will be evaluated in terms of the predictive accuracy (Relative Brier Score). The best set of intervals (that may not be those obtained in the first step) for each TN is selected based on predictive accuracy. When a TN has as parents other Temporal Nodes (an example of this situation is illustrated in Figure 4), the configurations of the parent nodes are not initially known. So, in order to solve this problem, the intervals are selected sequentially in a top-down fashion according to the TNBN structure.

The algorithm then iterates between the structure learning and the interval learning. However, for the experiments presented in this work, we present the results of the first iteration.

5 Experiments

5.1 Data and preprocessing

Data was obtained from the HIV Stanford Database (HIVDB) [19]. The isolates in the HIV Drug Resistance Database were obtained from longitudinal treatment profiles reporting the evolution of mutations in individual sequences.

In total data from 2373 patients with subtype B was retrieved. We choose to work with this subtype because it is the most common in America [10], our geographical region of interest. For each patient data retrieved contains a history consisting of a variable number of studies. Information regarding each study consists of a treatment or cocktail of drugs administered to the patient, how long the treatment lasted in weeks, and the list of more frequent mutations in the viral population within the host at the time when the treatment was suspended (changed for a different treatment). An example of the data is presented in Table 1.

Table 1. An example of the data. It presents two patients P_1 with 3 temporal studies, and P_2 with two temporal studies.

| Patient | Treatment | List of Mutations | Weeks |
|---------|---------------|-------------------|-------|
| P_1 | LPV, FPV, RTV | L63P, L10I | 15 |
| | | V77I | 30 |
| | | I62V | 10 |
| P_2 | NFV, RTV, SQV | L10I | 25 |
| | | V77I | 45 |

For applying the learning algorithm for TNBN the data presented in Table 1 is transformed into a table similar to the one presented in Table 2. Here, each column represents a drug or mutation, each row represents a patient case, for the drugs the values are USED or NOT USED, and for the drugs the values are: APPEAR with the number of weeks that mutation appeared the first time or Default, this is when mutation did not appear in that case. The ordering provided to the K2 algorithm is: first the antiretrovirals ordered by frequency, then the mutations ordered by frequency.

The number of studies available varies from 1 to 10 studies per patient history. Since we are interested in temporal evolution of the mutational networks, we filtered those patients having less than 2 studies, with 973 patients outliving the conditional.

Antiretrovirals are usually classified according to the enzyme that they target. We focus on protease as this is the smallest of the major enzymes in terms of number of aminoacids. There exist 9 protease inhibitors (PI), namely: Amprenavir (APV), Atazanavir (ATV), Darunavir (DRV), Lopinavir (LPV), Indinavir (IDV), Nelfinavir (NFV), Ritonavir (RTV), Tripanavir (TPV) and Saquinavir

Table 2. An example of the data used to learn the TNBN model. Each row represents a patient. Each column represents either a drug, used or not, or a mutation that appeared or not.

| Drug-1 | Drug-2 | ... | Mutation-1 | Mutation-2 | ... |
|--------|----------|-----|-------------|-------------|-----|
| USED | NOT USED | | (APPEAR) 30 | Default | |
| USED | NOT USED | | (APPEAR) 40 | Default | |
| USED | USED | | Default | (APPEAR) 80 | |

(SQV). All 9 PIs will be considered during the experiments. Figure 2 presents the histogram of the administration of the different PIs in the dataset. Data from HIVDB originates from different studies and in some cases is incomplete. In this sense, Figure 2 evidences a small portion of studies only reporting the administration of a PI, but the specific antiretroviral is missing. Also there is slightly bigger portion reporting *Unknown* as the drug used. The way we handle this cases was, if the patient case only contained *Unknown* or *None* that case was removed. However, if the case contained other drug (apart from *Unknown*), that information was used for the model.

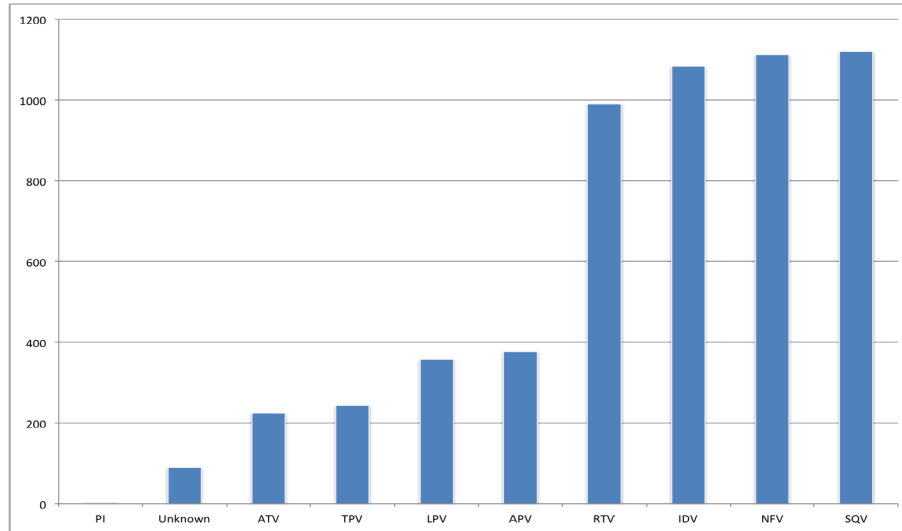


Fig. 2. Histogram of the protease inhibitors administration in the full dataset including all 2373 patients.

We still have to define a target set of mutations of interest. Figure 3 presents the histogram of mutations as they appear in the dataset. 733 different mutations appeared at least once in the data. But as evident from the histogram, most

mutations are rare presenting low frequency. Indeed, only few mutations exhibit high frequency. Mutation inclusion criteria is detailed below.

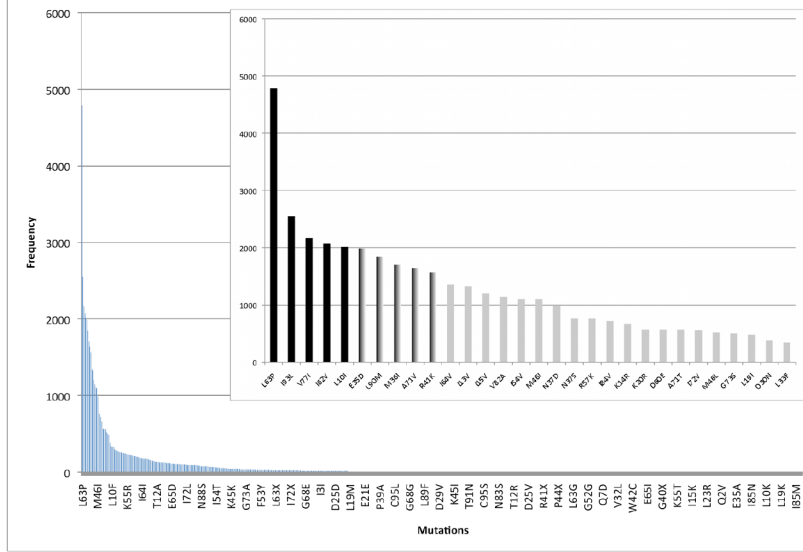


Fig. 3. A group of mutations and their frequency in the full dataset including all 2373 patients. The higher frequency end of the spectrum is zoomed. Mutations used for the first experiment are shown in black, and the completing set for the second experiment are indicated in shaded black.

5.2 Model evaluation

The learning algorithm arrives to local maxima, and thus is influenced by initial parameterization. For our experiments, the number of initial intervals for each node was allowed to vary from 2 to 4 and Equal Width Discretization [13] was used to initialize those intervals. Since it does not exist a gold standard or a reference TNBN, in evaluating the model's performance three indirect measures were used: the relative Brier score (RBS), the relative time error and the total number of intervals in the model. From the 973 total patients 80% of them were used for learning and the rest for evaluation.

The Brier Score is a measure of the predictive accuracy of the network, is defined as

$$BS = \frac{1}{n} \sum_{i=1}^n (1 - P_i)^2$$

where P_i is the marginal posterior probability of the correct value of each node given the evidence, this applies for all the selected nodes, n , of the TNBN. The

RBS is defined as:

$$RBS \text{ (in \%)} = (1 - BS) \times 100$$

For each case of the data (a row in table 1), the RBS is obtained by instantiating a random subset of variables in the model, predicting the unseen variables, and obtaining the RBS for these predictions. The relative time error is a measure to evaluate how far the real events are from the intervals and it is defined as the difference between the real event and the middle point of the interval divided by the range of the temporal node. The range of the node is the difference between the maximum and the minimum values of the intervals in a temporal node. Finally, the number of intervals is defined as the total number of intervals learned across all variables, this is a rough estimate of the complexity of the network and a low number of intervals is a desirable property for simplicity of the model. The best model would afford a high RBS, a low time error and a low complexity (low number of intervals). The technical performance of the model reflects its predictive accuracy and complexity, but should not be confused with the biological/physiological plausibility of the model.

5.3 Results

Two experiments have been carried out. The first experiment with a smaller model aims to assess the capability of TNBN for capturing known relations and thus providing a qualitative validation of the approach. The second, with a more complete model is aimed at uncovering the more common existing mutational networks and capturing the temporal aspect of the network formation.

In the first experiment, only mutations with more than 2000 counts were used: L63P, I93L, V77I, I62V and L10I. For tractability, in the second experiment only those mutations appearing more than 1500 times were included: L63P, I93L, V77I, I62V, L10I, E35D, L90M, M36I, A71V and R41K.

Table 3. Evaluation of the models in terms of RBS, relative time error (in percentage) and number of intervals.

| Experiment | Initial intervals | RBS | Relative Time error | Number of total intervals |
|------------|-------------------|-------------|---------------------|---------------------------|
| 1 | 2 | 89.8 | 13.0 | 17 |
| | 3 | 88.3 | 13.6 | 20 |
| | 4 | 88.5 | 13.9 | 19 |
| 2 | 2 | 87.3 | 15.0 | 30 |
| | 3 | 88.5 | 14.7 | 31 |
| | 4 | 87.5 | 15.9 | 35 |

Table 3 summarizes the results for the two experiments. Figure 4 illustrates the TNBN of the first experiment exhibiting the best scores. The figure represents the network, the intervals and the prior probabilities obtained for each TN.

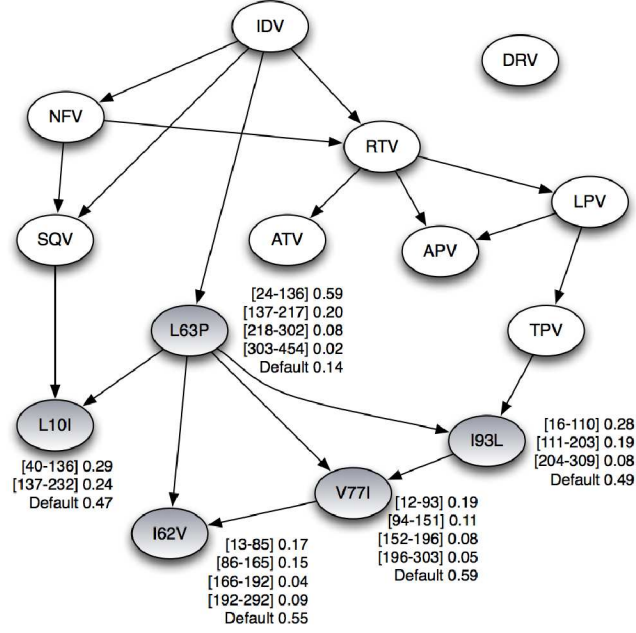


Fig. 4. A learned TNBN with 9 Protease Inhibitors and 5 mutations that appear frequently. Drugs are indicated in white bubbles and mutations in gray bubbles. Learned time intervals and their probabilities for the TNs are indicated beside the bubble as is standard in the representation of TNBNs.

Exploring the model reveals that RTV has arcs linking it with IDV, NFV, ATV, APV and LPV. This important relationship of RTV with other medicaments is explained due to the fact that the Ritonavir drug has been proved to boost the effect of other PIs, and therefore most of the times it is administered in combination with other drugs. The link between SQV and L10I was also already known to clinicians [12] and our model has also been successful in uncovering it. The observation of these known fact boosts our confidence that the model was meaningful; but can the model reveal new knowledge? The DRV node in the model is isolated because in the data, was never given as part of a first treatment for any of the patients. This is perhaps because DRV is a relatively new drug. L63P is an extremely common mutation in the viral genotype coding the protease as revealed by the histogram in Figure 3. However, “when” this mutation appears in the evolving virus remained mysterious. Our model suggests that most times this mutation tends to appear early in time, and that its probability to appear decreases over it.

Figure 5 illustrates the best TNBN model instantiation in terms of higher RBS for the second experiment. Most of the arcs from the smaller model were retained. Only the relation linking I62V and V77I and TPV with I93L are drop. Moreover, only two new arcs from SVQ and TPV to L63P appear among previ-

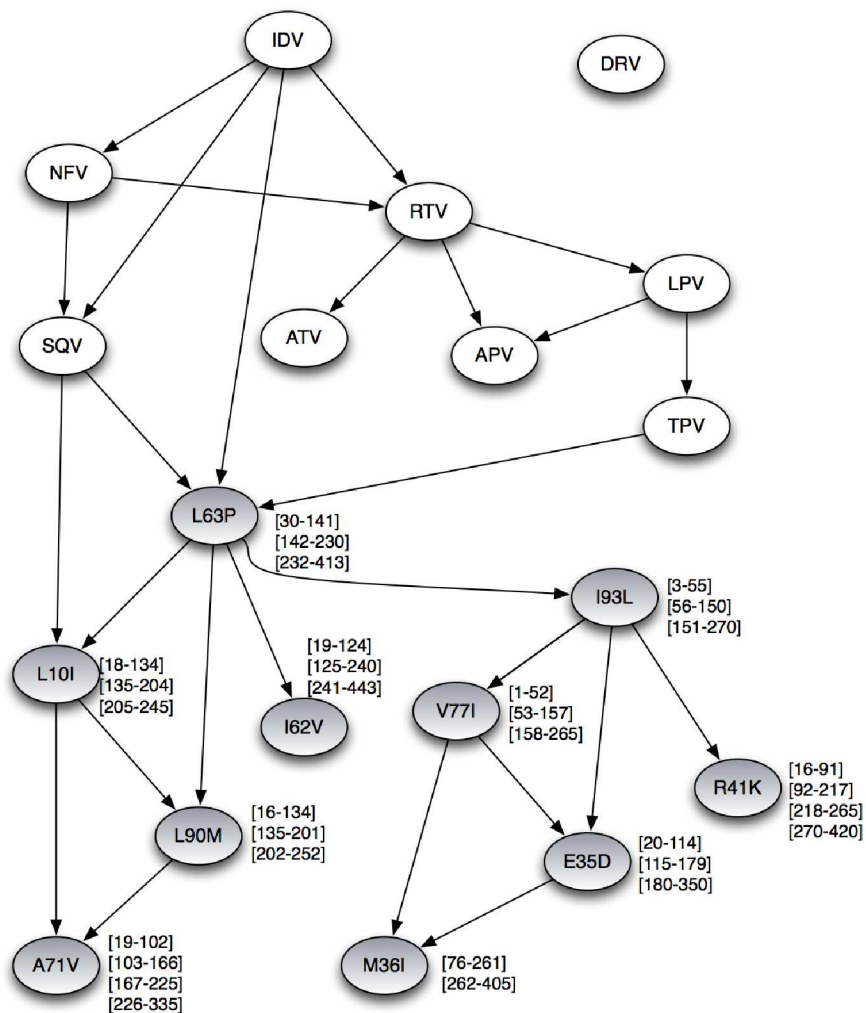


Fig. 5. A learned TNBN with 9 Protease Inhibitors and 10 mutations that appear frequently. Drugs are indicated in white bubbles and mutations in gray bubbles. Learned time intervals for the TNs are indicated beside the bubble. Probabilities and the Default state for the TNs are hidden for readability.

ously considered elements. This small variation among the two models is a good indicator of the robustness of the modeling approach. From this more complete model the prevalence of mutation L63P is evident from its relation to most drugs. It can then diverge to other mutations. There are two possible explanations for this observation; either the frequency of appearance of L63P is biasing the formation of the associations in the model -L63P almost doubles in frequency that of the following most common mutation-, or L63P is a key mutation to unleash

others. Additionally, the local neighborhoods in the graph clearly reveal two mutational networks;

* L63P, I62V, L10I, L90M and A71V

* I93L, V77I, M36I, E35D and R41K

We are not aware of these mutational networks to have previously been reported. Moreover, the TNBN further reveals the temporal sequence of mutation appearance.

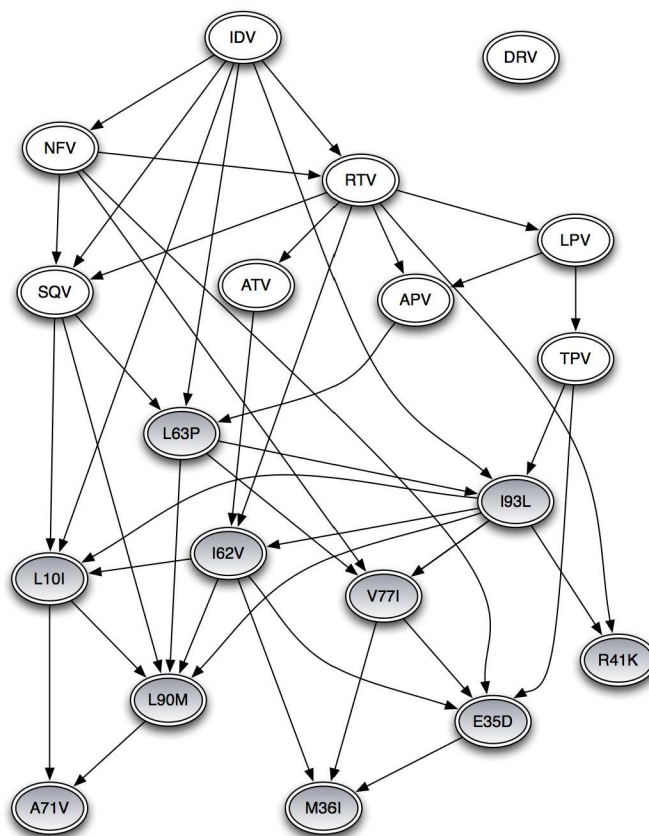


Fig. 6. A learned *static* BN with 9 Protease Inhibitors and 10 mutations that appear frequently. Drugs are indicated in white bubbles, their states are Used or Not Used. Mutations are presented in gray bubbles, their states are Appeared or Not Appeared. States and probabilities are hidden for readability.

As a final exercise, the temporal information of the mutations has been removed. Hence, the states for both the drugs and the mutations were: Appear

or Not Appeared. We used the same ordering as in the previous experiments and applied the K2 learning algorithm. The static BN learned is presented in Figure 6. The number of arcs increased. More importantly, most of the arcs obtained with the TNBN remained. Using temporal information in this particular case yields a simpler model. Further analysis of these experiments is needed, to dilucidate whether this is always the case. This last exercise while interesting computationally, is far from construct validating none of the models. In this sense we are of course short from being able to determine which one is more correct, even though intuitively the simpler TNBN seems more adequate by Occam’s razor.

6 Conclusions

By using a TNBN we have been able to unveil the two more common mutational networks present in HIV evolution as response to pharmacological selective pressure, and we believe these to be previously unreported. Our model has been successful in capturing relationships between mutations and protease inhibitors critically incorporating temporal information. These results are encouraging, presenting the model as an effective tool to explain how mutations interact with each other and providing some leverage for the clinicians in interpreting clinical tests. In this sense, the success of the second model still raises more questions. For example, why are ATV and APV not related to any mutation? This demands further investigation.

Models such as ours are an initial step to facilitate treatment planning. If a certain mutation occurring early in a mutational network is observed during a sequentiation, one would expect the other mutations in the network to follow. Knowing the likely appearing of subsequent mutations gives the therapist an edge in determining the appropriate antiretroviral regimen.

Future work plans to use TNBNs models to unfold mutational networks in the Reverse Transcriptase, another important enzyme of HIV. Technically, robustness of the model may be objectively assessed by means of bootstrap [8] to check which substructures remain in the model across different subsets of data. Finally, formal validation of the approach is still pending.

Acknowledgements

El trabajo que ha dado lugar a este invento ha recibido apoyo del CONACYT y de la Comunidad Europea a través del FONCICYT en virtud del contrato de asignación de recursos/contrato de subvención n 95185. The first author is supported by a grant 234507 from CONACYT.

References

1. Arroyo-Figueroa, G., Sucar, L.E.: A temporal Bayesian network for diagnosis and prediction. In: Proceedings of the 15th UAI Conference. pp. 13–22. Stockholm, Sweden (1999)

2. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J.: Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proceedings of the National Academy of Sciences of the United States of America* 99(12), 8271–8276 (2002)
3. Chausa, P., Cáceres, C., Sacchi, L., León, A., García, F., Bellazzi, R., Gómez, E.: Temporal Data Mining of HIV Registries: Results from a 25 Years Follow-Up. *Artificial Intelligence in Medicine* pp. 56–60 (2009)
4. Cooper, G., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. *Machine learning* 9(4), 309–347 (1992)
5. Dagum, P., Galper, A., Horvitz, E.: Dynamic network models for forecasting. In: *Proceedings of the 8th Workshop UAI*. pp. 41–48. Stanford, California, USA (1992)
6. Draghici, S., Potter, R.B.: Predicting HIV drug resistance with neural networks. *Bioinformatics* 19(1), 98–107 (2003)
7. Freeman, S., Herron, J., Payton, M.: *Evolutionary analysis*. Prentice Hall Upper Saddle River, NJ: (1998)
8. Friedman, N., Goldszmidt, M., Wyner, A.: Data analysis with Bayesian networks: A bootstrap approach. In: *Proceedings of the 15th UAI Conference*. pp. 206–215. Stockholm, Sweden (1999)
9. Galán, S., Arroyo-Figueroa, G., Díez, F., Sucar, L.: Comparison of two types of event Bayesian networks: A case study. *Applied Artificial Intelligence* 21(3), 185 (2007)
10. Hemelaars, J., Gouws, E., Ghys, P.D., Osmanov, S.: Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 20, W13–W23 (2006)
11. Hernandez-Leal, P., Sucar, L.E., Gonzalez, J.A.: Learning temporal nodes Bayesian networks. In: *The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*. Palm Beach, Florida, USA (2011)
12. Johnson, V., Brun-Vézinet, F., Clotet, B., Günthard, H., Kuritzkes, D., Pillay, D., Schapiro, J., Richman, D.: Update of the Drug Resistance Mutations in HIV-1: December 2010. *Topics in HIV medicine* 17, 138–145 (2010)
13. Liu, H., Hussain, F., Tan, C., Dash, M.: Discretization: An enabling technique. *Data Mining and Knowledge Discovery* 6(4), 393–423 (2002)
14. Neapolitan, R.: *Learning Bayesian networks*. Pearson Prentice Hall (2004)
15. Pang, B., Zhang, D., Li, N., Wang, K.: Computerized tongue diagnosis based on Bayesian networks. *Biomedical Engineering, IEEE Transactions on* 51(10), 1803–1810 (2004)
16. Pearl, J.: *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann (1988)
17. Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C.: The causes and consequences of HIV evolution. *Nature Reviews Genetics* 5(1), 52–61 (2004)
18. Ramirez, J., Cook, D., Peterson, L., Peterson, D.: Temporal pattern discovery in course-of-disease data. *Engineering in Medicine and Biology Magazine, IEEE* 19(4), 63–71 (2000)
19. Rhee, S., Gonzales, M., Kantor, R., Betts, B., Ravela, J., Shafer, R.: Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research* 31(1), 298–303 (2003)
20. Rodin, A., Boerwinkle, E.: Mining genetic epidemiology data with Bayesian networks: Bayesian networks and example application (plasma apoE levels). *Bioinformatics* 21(15), 3273–3278 (2005)

Bayesian data analytic knowledge bases for genetic association studies

P. Sarkozy¹, P. Marx¹, A. Millinghoffer¹, G. Varga², A. Szekely², Zs. Nemoda³, Zs. Demetrovics², M. Sasvari-Szekely³, P. Antal¹

(1)Department of Measurement and Information Systems, Budapest University of Technology and Economics

(2)Institute of Psychology, Eötvös Loránd University, Budapest, Hungary

(3)Institute of Medical Chemistry, Molecular Biology and Pathobiochemistry, Semmelweis University, Budapest, Hungary

Abstract. Bayesian methods and Bayesian networks are increasingly popular in genetic association studies. We discuss the application of Bayesian networks to give a detailed characterization of relevance relations and their application in case of multiple outcome variables. These global properties of the relevance relations are investigated in the Bayesian statistical framework using a joint model, thus we can generate a coherent uncertainty measure for the results without post hoc corrections. We show the usefulness of the syntactic aggregation of the a posteriori distributions over the relevant variable sets, which allows the examination of the most relevant variables, variable pairs, and larger subsets. We present these methods as precursors for a unified framework of Bayesian data analytic knowledge bases describing the results of multiple Bayesian analysis of relevance. Concepts are demonstrated in the genetics of trait impulsivity.

1 Introduction

Genetic association studies face many challenges such as the poor description of phenotypes, presence of population confounding, effects of life style and environment, the seemingly non-functional nature of the factors found, the weak effect strength of the factors (“missing heritability”), but the most profound is the rapid increase of the number of potential predictors, which manifests itself as “the multiple hypothesis testing problem” in the frequentist framework. In response to this limit, more intensive usage of computational resources and background knowledge became central issues in biomedicine. In genetic association studies such approaches have emerged in various contexts to cope with the relative scarcity of the data such as the pooling of datasets in meta-analysis, pooling of the results in ad hoc repositories and knowledge bases, and the use of computation-intensive statistical approaches such as permutation testing, bootstrap, and Bayesian statistics.

In the paper we present elements of a Bayesian, global relevance analysis and show their application in the probabilistic knowledge fusion research direction in the following aspects:

1. *Partial (strong) relevance* We can infer the a posteriori probability the k variables are jointly strongly relevant for a given outcome potentially with further unspecified variables.
2. *Type of relevance* We can infer posteriors for various types of relevance, e.g. strong relevance vs. association.
3. *Multi-target relevance* We can infer posteriors for strong relevance w.r.t. multiple outcome variables.

The advantages of Bayesian networks (BN) for representing global dependency maps and relevance relations are well-known, but their application was hindered in high-dimensional tasks by their high computational and sample (statistical) complexity. Motivated by this problem we proposed a Bayesian approach to the feature subset selection (FSS) problem and proposed the use of partial relevance and multi-target relevance [2]. In this paper we extend this approach by inferring and comparing posteriors for various subtypes of pairwise dependencies, such as association and strong relevance.

First in Section 2 we overview Bayesian network based concepts of relevance and earlier applications. In Section 3 we discuss the Bayesian approach to FSS, particularly the main assumption of its popular conditional version, which makes it different from the general, domain model based approaches, and summarize the Stochastic Search Variable Selection (SSVS), which is one of our evaluation methods. Then in Section 4 we overview earlier Bayesian network based methods in the Bayesian framework to analyze relevance and summarize our approach. Section 5 and Section 6 contains the results in impulsivity research and its discussion.

2 Bayesian network representation of relevance

There are many association analysis methods with different biases, advantages and disadvantages w.r.t the number of variables, sample size, quality and completeness of the data, loss function, time, and available computational resources. Thus an important point of reference is an asymptotic, loss-free, algorithm-free probabilistic concept of relevance, the Markov Blanket Set (MBS) [29]. It was connected to the Bayesian networks (BN), which became a central tool for the graphical representation of dependencies and optionally causation [24]. In the feature (attribute) learning context related univariate concepts of relevance, strong and weak relevance

was introduced [20]. To bridge the gap between the linear cardinality of the Markov blanket membership (MBM)s and exponential cardinality MBSs, we introduced the concept of partial (multivariate, strong) relevance (k-MBS) with scalable, intermediate polynomial cardinalities [2].

Because in our application domain the outcome variables are semantically related, we use the following acasual subtypes of relevance, which are derived from the combinations of {causal,confounded,conditional}, and {direct,indirect} relations and their aggregates, see Table 1 (for a causal interpretation under the Causal Markov Assumption, see e.g. [24]).

Table 1. Graphical model based definition of types of relevances and associations.

| Relation | Abbreviation | Graphical |
|-------------------------------|--------------|--|
| Direct causal relevance | DCR(X,Y) | There is an edge between X and Y |
| Transitive causal relevance | TCR(X,Y) | There is directed path between X and Y |
| Confounded relevance | ConfR(X,Y) | X and Y have Common ancestor |
| (Pairwise) Association | A | DCR or TCR or ConfR |
| Pure interactionist relevance | PIR(X,Y) | X and Y have common child |
| Strong relevance | SR(X,Y) | PIR or DCR |

The ordering of relations in Table 1 indicates certain ontological, and practical aspects, but a hierarchy or ranking is problematic, because for example the standard concept of pairwise association (A) is narrower than strong relevance (it does not include Pure Interactionist Relevance). Further extension of relevance is possible, if there are multiple possible target variables \mathbf{Y} which have to be examined together, thus we proposed the the concept of multi-target relevance [2].

The Markov Blanket Set and the Bayesian network representation induced many research direction in feature learning, in the feature subset selection problem, and in genetic association studies [11, 21, 34, 16, 1, 18, 36]. Because of the high computational complexity and particularly because of the high sample (statistical) complexity of learning complete Bayesian network models w.r.t number of variables these “local” approaches limit their scope, and focus on the identification of strongly relevant variables, and possibly their interaction and causal structure. Thus global and detailed characterization of relevance relations is not available. However as we will show the Bayesian statistical framework provides a normative solution for the high sample complexity and for medium sized problems with hundreds of variables the computational complexity is manageable using high-throughput and high-performance computing resources.

3 The conditional Bayesian approaches and the SSVS

Bayesian methods are more and more popular in genetic association studies, and one of their advantage is their principled approach to model complexity and number of variables in case of relatively small sample size [6]. The infamous correction for multiple hypothesis testing with frequently ad hoc management - causing loss of significance and power - manifests itself in the Bayesian framework as a normative and inherent property, resulting in a more flat posterior for more complex models.

In the feature learning context a popular choice is the conditional Bayesian approach, which assumes independent beliefs corresponding to the modeling of the dependence of the output variable Y on X (i.e., without modeling the overall domain) [14]. Practically the conditional approach models the conditional distribution of Y given \mathbf{X} using a parametric model class $S, \boldsymbol{\theta}$ as $p(Y = 1|X = x, S, \boldsymbol{\theta})$, for example using linear regression, logistic regression or multilayer perceptrons. The domain model based approach models the joint distribution of Y, \mathbf{X} using a parametric model class $S, \boldsymbol{\theta}$ as $p(Y, \mathbf{X}|S, \boldsymbol{\theta})$, for example using Bayesian networks. In both cases using the posterior over model structure $p(S, \boldsymbol{\theta}|D_N)$ given a data set D_N we can induce a posterior for the relevance of a feature X_i and for the subset of features \mathbf{X}' . A fundamental difference between the conditional and domain model based approach is that in the conditional approach the presence of a variable in the model can not be interpreted as strong relevance (e.g. a highly predictive, but only weakly relevant factor can be present in the conditional model, if it is strongly associated through multiple paths, as we do not model the dependencies between the factors).

An early Bayesian conditional approach, the Stochastic Search Variable Selection puts the regression problem in a Bayesian statistical framework. This approach considers submodels with subsets of the predictor variables and estimates the a posteriori probability of the inclusion of a predictor and its corresponding strength parameters [15]. Bayesian variable selection method is based on assuming a normal prior distribution on the regression parameters. The variance of the distribution usually is a constant, but we can extend the model by estimating the variance as in case of SSVS. If we estimate the variance of normal prior, it helps tuning the parameters, because in the regression model the coefficient depends on the variance. In a heterogeneous problem, the variance can be set differently for all regression variables.

Other Bayesian conditional methods e.g. using logistic regression or multilayer perceptrons, are widely used in biomedicine and in GASs (e.g., see [3, 27, 6, 31, 28, 35, 32, 12]). Although the conditional approach is capable for multivariate analysis including interactions, the domain model based approach allows better characterization of both local and global dependencies.

4 Bayesian analysis of relevance using Bayesian networks

The local “causal” discovery methods limit their scope to the strongly relevant variables to reduce the high computational complexity and particularly the high sample (statistical) complexity of learning complete Bayesian network models, i.e. to avoid the learning of a global and detailed characterization of relevance relations [1]. However the Bayesian statistical framework provides a normative solution for the high sample complexity and for medium sized problems with hundreds of variables the computational complexity is manageable using high-throughput and high-performance computing resources. Thus the BN based Bayesian approach can ensure global, potentially causal characterization of the dependencies and normative characterization of weakly significant results.

The Bayesian inference over structural properties of Bayesian networks was proposed in [7, 9]. In [22], Madigan et al. proposed a Markov Chain Monte Carlo (MCMC) scheme to approximate such Bayesian inference. In [13], Friedman et al. reported an MCMC scheme over the space of orderings. In [19], Koivisto et al. reported a method to perform exact full Bayesian inference over modular features. An ad hoc randomized approach were reported in [30]. For the application of Bayesian networks in the Bayesian framework we reported specialized ordering MCMC methods to efficiently estimate posteriors over structural model properties, particularly over Markov Blanket Graphs (MBG) (the ordering-conditional posterior of an MBG can be computed in polynomial time, which can be exploited in ordering-MCMC methods [4]). Based on these concepts we proposed a Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA), which estimates posteriors of hierarchic, interrelated hypotheses, e.g. for partial strong relevance for all subsets. Partial strong relevance is particularly useful, because it defines an embedded hypotheses space with varying complexity, i.e. sets of k predictors that are strongly relevant [2].

The posteriors for the hierarchic, interrelated hypotheses of the BN-BMLA methodology are estimated in a two-step process to support post-

hoc analysis. First we estimate posteriors over the MBS, MBG, and for the pairwise relations in Table 1 for the target variables. In the second phase we use these posteriors as a probabilistic knowledge-base to estimate various posteriors and discover interesting and significantly confirmed hypotheses. In the first phase we applied MCMCM method over the Bayesian network structures (i.e. over directed acyclic graphs, DAGs) without limiting the maximum number of parents. We used both the Cooper-Herskovits (CH) and the observationally equivalent BDeu parameter priors with various virtual sample sizes (VSS=1,10,50,100), but from the point of view of biomedical relevance we found that the theoretically preferable BDeu prior is more sensitive to "small-sample" anomalies, thus we report results for the CH and VSS=1 setting. The structure prior was uniform. The length of the burn-in and MC simulation is 10^6 and 5×10^6 , the probability of the DAG operators is uniform [8]. In the second step we computed offline the k-MBS posterior values from the MBS posterior, the posteriors over the types of the dependency relations in Table 1, and the posteriors for multi-target relevance.

5 Results

Impulsivity or impulsiveness is a personality trait defined as a predisposition toward rapid, unplanned reactions. We investigated a combined set of serotonergic (HTR1A-1019 C/G, HTR1B 1997 C/G, 5-HTTLPR in SLC6A4 gene) and dopaminergic (COMT Val158Met, DRD4 48bp VNTR, DRD2/ANKK1 Taq A1) polymorphisms. The sample size was 561, which included only complete records from a preliminary dataset of a larger study. The impulsivity phenotype was measured by the Hungarian version of the Barratt Impulsivity Scale (BIS-11) originally published by Patton and colleagues [23]. The instrument consists of 30 items, scored on a four point scale. The three main impulsivity factors are: Cognitive, Motoric, and Nonplanning impulsiveness. The total score is the sum of all items.

To cope with multiple predictors with potentially weak effects we applied the stochastic search variable selection method (SSVS) to the genotypes, Sex and Age data, while normalizing the scale targets. We used the SSVS for quantile regression implementation in the MCMCpack package [25] in R [26]. We ran the algorithm for the different scale targets with the same parameter setting to get comparable results. We set the shape parameters of the beta distribution to 10 and 2. We ran 100000 iterations and 10000 for the burn-in period. Two of the predictor variables was

found significant in case of all target variables (Fig. 1). These two variables (HTR1B and DRD4) have significantly higher marginal inclusion probability. The regression coefficients with the highest absolute value belongs also to these two predictors HTR1B and DRD4 (Fig. 5).

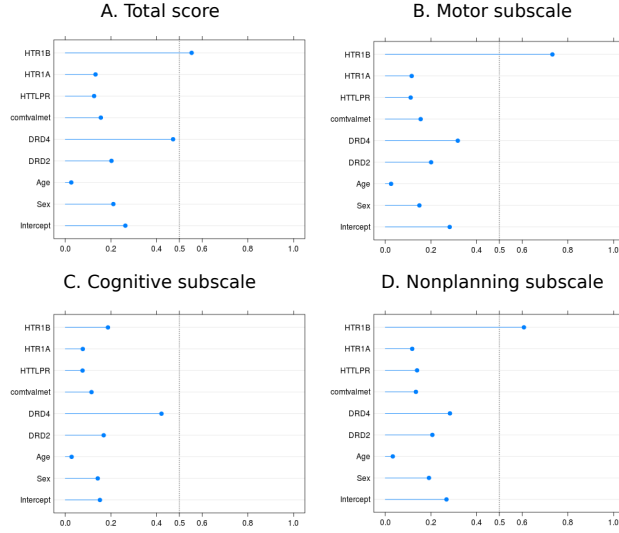


Fig. 1. Marginal posterior probability for all predictors. The posterior probabilities (X axis) for the total score (A), for the motor subscale (B), for the cognitive subscale (C) and for the nonplanning subscale (D).

To investigate the interactions; types of the relevance of the predictors; and their relevance in a joint analysis of multiple target variables we applied the BN-BMLA method. It was applied for the 5-HTTLPR genotypes and five other grouped genotype categories, as well as Sex, and Age. Outcome (target) variable was the BIS Total score or the scale variables separately and jointly. The scale variables and Age were discretized into three bins with uniform univariate frequencies.

The identification of an overall domain model was not possible, because considerable uncertainty remained at the level of full multivariate analysis (see Fig. 5). Therefore we computed the aggregate posterior probabilities for variables, pairs of variables, and triplets of variables. Fig. 5 reports a comparative overview of peakness of the posteriors for uni-, bi-, and trivariate partial strong multi-target relevance. It shows that DRD4 is strongly relevant (with 0.575 posterior probability), the DRD4 and HTR1B pair is among the strongly relevant variables (with 0.251 pos-

Table 2. The regression coefficients for the predictors in case of the three subscales and the total score

| Predictors | Total score | Motor subscale | Cognitive subscale | Nonplanning subscale |
|------------|-------------|----------------|--------------------|----------------------|
| Intercept | -5.821e-03 | -2.730e-02 | 0.01 | -0.008 |
| Sex | 2.264e-02 | 4.126e-03 | 0.022 | 0.0151 |
| Age | -4.644e-05 | 6.403e-05 | -0.0006 | 0.0003 |
| DRD2 | -6.593e-03 | -1.093e-02 | -0.016 | -0.0069 |
| DRD4 | -1.074e-01 | -4.822e-02 | -0.09 | -0.04 |
| COMT | -1.122e-02 | -8.690e-03 | -0.012 | -0.0038 |
| 5-HTTLPR | 4.721e-03 | 1.588e-03 | -0.0045 | 0.006 |
| HTR1A | 4.670e-03 | -5.480e-04 | -0.004 | 0.0002 |
| HTR1B | -1.391e-01 | -2.003e-01 | -0.023 | -0.16 |

terior probability). Posterior probabilities of three member variable sets showed no marked features. The standard errors of the estimated posteriors are below 0.001. Fig. 5 shows the most probable MBGs with posterior larger than 0.001 (the width of the edges indicate their aggregate posteriors).

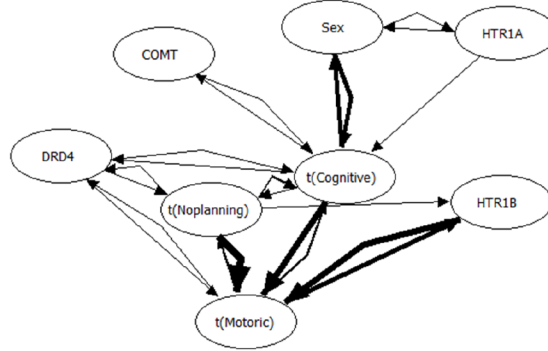


Fig. 2. The most probable Markov Blanket Graphs with posterior larger than 0.001 (the edge posteriors are indicated by their width).

We also computed the posteriors for the types of the relevance of the predictors in Table 1. As expected it provides a useful, detailed characterization, e.g. in case of DRD2 the a posteriori probability of association between DRD2 and bisTotal is 0.4926, but its strong relevance is only 0.0219. Note that such interpretation and decomposition is not available with SSVS.

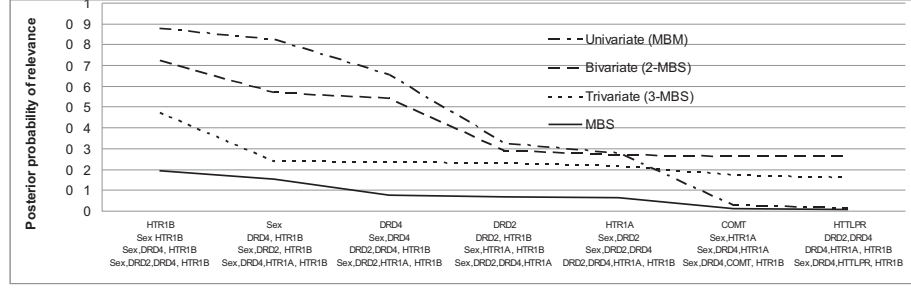


Fig. 3. The most probable uni-, bi-, and tri-variate subsets with decreasing multi-target relevance.

Finally we performed a refined analysis for multiple outcomes and applied the BN-BMLA method for the three subscales separately, also for the BIS Total score, and also jointly to compute the multi-target relevance. This confirmed that HTR1B is strongly relevant for the motoric and noplanning subscales, DRD4 has somewhat weaker, but similar multiple effect, but interestingly Comt and HTR1A is strongly relevant only for the cognitive subscale.

6 Discussion

The applied methods (SSVS, BN-BMLA) gave similar results. Both confirmed that the serotonergic and dopaminergic polymorphisms affect the trait impulsivity scores, specifically DRD4 and HTR1B. Furthermore BN-BMLA provided a coherent characterization of the system of dependencies and a detailed picture of the genetic background of the subscales which makes it a promising option in genetic studies. The Bayesian network based analysis confirmed association of DRD4 with BIS Total, moreover it was also weakly linked to all three subscales. With respect to this variable set the effect of DRD4 was direct, i.e. it was strongly relevant. HTR1B showed marked effects only towards the BIS Total score and towards the Motor subscale. The analysis showed that there was no statistical interaction between these two variables, which was confirmed by posterior decomposition analysis [2].

We analyzed partial multivariate strong relevances, because the Bayesian statistical framework allows the calculation of posteriors for the strong relevance of variables, pairs of variables, triplets of variables, etc. This is more flexible than the complete relevance patterns of all the variables, because it allows the selection of appropriate level of complexity of hypothe-

ses. As shown in Fig 5 the relevance of such subsets of variables exhibit differently peaked distributions, which are in close correspondence with feature complexity. These results indicated weak associations for HTR1A and Sex.

These preliminary results and other applications indicate that Bayesian networks offers a rich language for the detailed representation of types of relevance, including causal, acausal, and multi-target aspects. Additionally Bayesian statistics offers an automated and normative solution for the multiple hypothesis testing problem, thus using high-throughput and high-performance computing resources posteriors for global(!), detailed characterization of relevance relations can be estimated in medium sized problems (i.e. for hundreds of variables). This Bayesian statistical, global relevance analysis extends the scope of local “causal” discovery methods and because of the direct interpretation of Bayesian posteriors contrary to p-values from the frequentist approach, it is an ideal candidate for creating probabilistic knowledge bases to support off-line meta-analysis and fusion of background knowledge.

The coherent characterization of the uncertainties over the detailed types of relevances offers the opportunity to interpret the results of a Bayesian GAS analysis as a “Bayesian data analytic knowledge base”. Currently we are working on techniques to allow the fusion of multiple Bayesian data analytic knowledge bases in related domains and support offline meta-analysis.

7 Acknowledgements

P.M. performed the SSVS based computations, P.S. performed the BN based computations and postprocessing. A.M. implemented the DAG-MCMC methods and algorithms for partial relevance, multiple relevance, and relevance types. A.Sz.,G.B. collected the data. M.S. performed the genotyping. P.A. designed the Bayesian analysis of partial relevance, multiple relevance, and relevance types. All the authors agree on contents of this paper. This work was supported by the NIH R03 TW007656 Fogarty International Research grant to Maria Sasvari-Szekely and the following Hungarian Scientific Research Funds: OTKA K81466 to Maria Sasvari-Szekely; OTKA-PD-76348 and NKTH TECH 08-A1/2-2008-0120 (Genagrid) to P. Antal. Anna Szekely and Peter Antal acknowledge the financial support of the János Bolyai Research Fellowship by the Hungarian Academy of Sciences. The authors thank Gabriella Kolmann for her technical assistance.

References

1. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research*, 11:171–284, 2010.
2. ANTAL, P., MILLINGHOFFER, A., HULLAM, G., SZALAI, C. and FALUS, A. 2008. A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction. JMLR Workshop and Conference Proceedings.
3. P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, 29:39–60, 2003.
4. P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.
5. D. J. Balding. A tutorial on statistical methods for population association studies. *Nature*, 7:781–91, 2006.
6. Stephens M. and Balding D.J. Bayesian statistical methods for genetic association studies. *Nature Review Genetics*, 10(10):681–690, 2009.
7. W. L. Buntine. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.
8. P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.
9. G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
10. COOPER, G. F. & HERSKOVITS, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309–347.
11. G. F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.
12. B. I. Fridley. Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology*, 33:27–37, 2009.
13. N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.
14. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
15. E. I. GEORGE, R. E. MCCULLOCH, 1993. Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association*, 88(423):881–889.
16. B. Han, M. Park, and X. Chen. A markov blanket-based method for detecting causal snps in gwas. *BMC Bioinformatics*, 11(3):5, 2010.
17. HULLAM, G., ANTAL, P., SZALAI, C. & FALUS, A. 2010. Evaluation of a Bayesian model-based approach in GA studies. JMLR Workshop and Conference Proceedings.
18. X. Jiang, M. M. Barmada, and S. Visweswaran. Identifying genetic interaction in genome-wide data using bayesian networks. *Genetic Epidemiology*, 34:575–581, 2010.
19. M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

20. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
21. D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
22. D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520, 1996.
23. PATTON, J. H., STANFORD, M. S. & BARRATT, E. S. 1995. Factor structure of the Barratt impulsiveness scale. *J Clin Psychol*, 51, 768–74.
24. PEARL, J. 2000. Causality: Models, Reasoning, and Inference, Cambridge University Press.
25. A. D. Martin, K. M. Quinn, J. H. Park, 2011. MCMCpack: Markov Chain Monte Carlo in R, *Journal of Statistical Software*, 42(9):1-21
26. R Development Core Team 2011, R: A Language and Environment for Statistical Computing,
27. C. Kooperberg and I. Ruczinski. Identifying interacting snps using monte carlo logic regression. *Genet Epidemiol*, 28(2):157–170, 2005.
28. M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2007.
29. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.
30. J.M. Pena, R. Nilsson, J. Björkgren, and J. Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232, 2007.
31. M. A. Province and I. B. Borecki. Gathering the gold dust: Methods for assessing the aggregate impact of small effect genes in genomic scans. In *Proc. of the Pacific Symposium on Biocomputing (PSB08)*, volume 13, pages 190–200, 2008.
32. B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate genes and quantitative traits. *PLoS Genetics*, 3(7):e114, 2007.
33. I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevance, filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.
34. Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.
35. Y. Zhang and J. S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173, 2007.
36. H. Xing, P. D. McDonagh, J. Bienkowska, T. Cashorali, K. Runge, R. E. Miller, D. DeCaprio and B. Church, R. Roubenoff, I. Khalil, and J. Carulli. Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Computational Biology*, 7(3):1001105, 2011.

Statistical Relational Learning for Clinical Domains

Jesse Jon Davis

Department of Computer Science
Katholieke Universiteit Leuven

Abstract. Machine learning has become an essential tool for analyzing biological and clinical data, but significant technical hurdles prevent it from fulfilling its promise. Standard algorithms make two key assumptions: the training data consist of independent examples and each example is described by a pre-defined set of attributes. Biomedical domains consist of complex, inter-related, structured data, such as patient clinical histories, molecular structures and protein-protein interaction information. The representation chosen to store the data often does not explicitly encode all the necessary features and relations for building an accurate model. For example, when analyzing a mammogram, a radiologist records many properties of each abnormality, but does not explicitly encode how quickly a mass grows, which is a crucial indicator of malignancy. This talk will describe an approach that automatically discovers unseen features and relations from data, which has advanced the state-of-the-art for machine classification of abnormalities on a mammogram. Presently most of the women identified for a possible malignancy on a mammogram are called back unnecessarily, with concomitant stress, procedure (additional imaging and/or biopsy) and expense. This research, which achieves superior performance compared to both previous machine learning approaches and radiologists, has demonstrated the potential to dramatically reduce this fraction without reducing the number of cancers correctly diagnosed.

Overview

Statistical relational learning (SRL) [5, 6], which combines first-order logic with probability, can model the complex, uncertain, structured data that characterizes clinical and biological domains. In these types of problems, the available data often does not contain all the necessary features and relations for building an accurate model. However, most SRL algorithms are constrained to use a pre-defined set of features and relations during learning. Requiring a domain expert to hand-craft all relevant features or relations necessary for a problem is a difficult and often infeasible task. Ideally, the learning algorithm should automatically discover and incorporate relevant features and relations.

The Score As You Use (SAYU) algorithm [1–4] is a general SRL framework for discovering new features and relations during learning. SAYU defines features

and relations as first-order logical rules and evaluates each one by building a new statistical model that incorporates it. If adding the new feature or relation improves the model’s predictive performance, then it is retained in the model. SAYU has been successfully applied to several tasks, including diagnosing breast cancer from structured mammography reports and predicting three-dimensional Quantitative Structure-Activity Relationships (3D-QSAR) for drug design.

Labeling an abnormality as benign or malignant from a structured mammography report is a challenging task for both radiologists and machines. Previous machine learning approaches to this problem are limited to using pre-defined features and ignore the relational nature of this task. However, a radiologist might include derived features which incorporate data about other abnormalities on the same mammogram or prior abnormalities in the decision process. SAYU, which can construct additional features that incorporate relational information, significantly outperformed radiologists, a hand-crafted Bayesian network system, standard Bayesian network structure learners and other SRL systems [2, 4] for this task.

References

1. Jesse Davis, Elizabeth Burnside, Inês Dutra, David Page, and Vítor Santos Costa. An integrated approach to learning Bayesian networks of rules. In *Proceedings of the 16th European Conference on Machine Learning*, pages 84–95. Springer, 2005.
2. Jesse Davis, Elizabeth Burnside, Inês C. Dutra, David Page, Raghu Ramakrishnan, Vítor Santos Costa, and Jude Shavlik. Learning a new view of a database: With an application to mammography. In Lise Getoor and Ben Taskar, editors, *An Introduction to Statistical Relational Learning*. MIT Press, 2007.
3. Jesse Davis, Vítor Santos Costa, Soumya Ray, and David Page. An integrated approach to feature construction and model building for drug activity prediction. In *Proceedings of the 24th International Conference on Machine Learning*, pages 217–224. ACM Press, 2007.
4. Jesse Davis, Irene Ong, Jan Struyf, Elizabeth Burnside, David Page, and Vítor Santos Costa. Change of representation for statistical relational learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2719–2726. AAAI Press, 2007.
5. Luc De Raedt, Paolo Frasconi, Kristian Kersting, and Stephen Muggleton, editors. *Probabilistic inductive logic programming: theory and applications*. Springer-Verlag, Berlin, Heidelberg, 2008.
6. Lise Getoor and Ben Taskar, editors. *An Introduction to Statistical Relational Learning*. MIT Press, 2007.

A Probabilistic Logic of Qualitative Time

Maarten van der Heijden^{1,2} and Peter J.F. Lucas²

¹ Institute for Computing and Information Sciences,
Radboud University Nijmegen, The Netherlands

² Department of Primary and Community Care,
Radboud University Nijmegen Medical Centre, The Netherlands
{m.vanderheijden, peterl}@cs.ru.nl

Abstract. Representing and reasoning over dynamic processes can benefit from using a qualitative representation of time when precise timing information is unavailable. Allen’s interval algebra provides a framework to do so. However, a more precise model can be obtained by modelling uncertainty in the process. In this paper we propose an extension of Allen’s algebra with uncertainty, using probability theory. Specifically, we use the expressive power of CP-logic, a probabilistic logic, to represent and reason with uncertain knowledge in combination with Allen’s interval algebra to reason qualitatively about time.

1 Introduction

In solving problems, one often has to take into account the time when a particular event has occurred or is expected to occur. Typically, the actual temporal details about when events have occurred are not available, or at least imprecise, whereas one is more certain about the actual order of the events. Medicine is a field where much of the information about patients has such a temporal, yet imprecise dimension. AI researchers have traditionally used Allen’s interval algebra [1] to model situations where there is much imprecision about the temporal evolution of events. Although Allen’s algebra supports qualitative reasoning about time, it does not allow expressing uncertainty about the qualitative, temporal relationships. Yet, uncertainty is a typical characteristic of many problems where precise temporal information is missing; medicine can again be taken as a prototypical domain for which this is true. Work by Shahar [2] clearly indicates the usefulness of Allen’s algebra for describing temporal events in medicine, and provides a use case in the form of temporal abstraction. For example, in hospital intensive care, interpreting the large amounts of (temporal) data becomes more manageable if we abstract from individual time points to a more qualitative representation. In this paper we aim to develop a happy marriage between Allen’s interval logic and uncertainty reasoning by making use of probabilistic logic as a unifying formalism.

Frameworks that combine logic and uncertainty have garnered quite some attention in the past few years. Specifically, various authors have proposed probabilistic logics, combining the expressive power of (a subset of) predicate logic representations with probabilistic uncertainty calculus. Examples of such temporal logics include probabilistic Horn abduction [3], Bayesian logic programs [4], and more recently ProbLog [5] and CP-logic [6], among others. Many of these probabilistic logics could serve as

the basis for an extension of Allen’s algebra; we argue that at a conceptual level CP-logic is already well aligned and is thus a natural choice. Allen [7] provides a temporal logic based on his interval algebra to model processes over time. We want to be able to reason about when certain events happen and how they relate to other events. It is then quite reasonable to take a causal viewpoint, as time and causality are closely related – causes precede their effects – and describing a temporal process as a causal mechanism seems an intuitive representation. CP-logic is short for ‘causal probabilistic logic’ and its semantics favours descriptions from a causal viewpoint, which meshes well with process descriptions of the kind you would want to specify in Allen’s logic. So if we are able to reason with Allen’s interval algebra within CP-logic, we obtain something that is conceptually pleasing while being more expressive than Allen’s logic.

Throughout the paper we will use a medical example to illustrate the developed concepts, because as mentioned, clinical medicine is a typical environment in which uncertainty and time play an important role. Specifically, we look at examples related to chronic obstructive pulmonary disease, COPD for short, and related lung diseases and complications. COPD is a progressive lung disease which is characterised by a mixture of chronic bronchitis and emphysema, leading to decreased respiratory capacity and potentially to respiratory failure and death. Although there are a number of causes, (tobacco) smoke is the most prevalent.

Because COPD is a progressive disease, its temporal development is quite important and even more so because of the occurrence of *exacerbation* events – a worsening of symptoms with possibly a large negative influence on health status. In modelling these kinds of situations many factors are uncertain. Often you do not know whether, for example, an exacerbation will occur, and even if you do you may not know when exactly. Allen’s algebra consists of qualitative relations that partially model the temporal uncertainty, yet Allen also recognised that modelling interesting processes that develop through time we need more than just temporal relations. The logic he proposed [7] combines the temporal relations with operators akin to predicate logic in order to state structural properties about the domain. Basically it is logic that provides the expressiveness to model things like causal connections, while the interval relations express temporal information and at least some of the uncertainty involved. However, the uncertainties of, for instance, predicting whether an exacerbation will occur given some observations of patient symptoms related in time, requires more extensive modelling capabilities. Advances in probabilistic logic provide us with tools that might help in modelling these uncertainties.

This paper is organised as follows. First, we go over some preliminaries, specifically, reviewing Allen’s temporal algebra somewhat more formally in Section 2.1 and 2.2, followed in Section 2.3 by a description of CP-logic, the probabilistic logic we use. Then in Section 3 we describe the extension to probabilistic temporal logic.

2 Preliminaries

2.1 Allen’s interval algebra

Allen’s algebra builds upon qualitative relations between time intervals. An interval implicitly refers to an event that takes place during that interval. When specifying an

event, the interval when the event happens is made explicit. To start, we will review the various possible relations between time intervals. Later we will discuss how such intervals, and their relationships, can be used to specify the evolution of events.

Two intervals can have a number of qualitative relations, some event can for example happen before another or overlap with it. The special treatment to relate processes for which exact timing information is unavailable, is offered by Allen's algebra. Allen's algebra defines 13 basic interval relations: *before*, *meets*, *overlaps*, *starts*, *during*, *finishes*, *equals* and their inverses. The inverse of a relation should be interpreted as the relation that holds when the intervals are interchanged, for example if interval i_1 is before i_2 , i_2 is after i_1 , thus *before* and *after* are each other's inverse. In Figure 1 the relations are shown graphically. They are mutually exclusive and complete in the sense that any two intervals can be assigned exactly one relation.

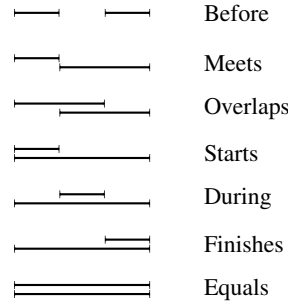


Fig. 1. Graphical representation of the seven basic relations that can hold between two time intervals. These relations and their inverse make up Allen's algebra.

Formally, we define intervals on a linearly ordered time line of points (\mathcal{T}, \leq) , which we take to be a subset of the set of the natural numbers \mathbb{N} . In the following we use the common abbreviations $<, \leq, =, \geq, >, \neq$ with their usual meaning to denote ordering and equality relations between elements of \mathcal{T} .

Definition 1. An interval I is defined as a pair of time points $I = [l, u)$, with $l, u \in \mathcal{T}$, using the convention of right-open intervals. We then define two special points $I^- = l = \inf I$ and $I^+ = u = \sup I$ to distinguish the start and end of I .

The following properties follow from the definitions given above:

Property 1: (trichotomy law) Only one of either $t < t'$, $t = t'$ or $t > t'$ holds.

Property 2: For each interval $I^- < I^+$.

Property 3: The number of intervals from \mathcal{T} is equal to $\binom{|\mathcal{T}|}{2}$

Property 1 follows from the linear order. Property 2 follows from the definition of intervals and their nonempty nature. Finally, property 3 follows from the fact that each choice of two time points constitutes an interval, and that the start of an interval is always less than the end. Thus, the number of possible intervals is equal to the number

of ordered pairs taken from the set \mathcal{T} , where one number is always less than or greater than the other number. We can now define relations on intervals.

Definition 2. Let \mathcal{I} be the set of all intervals of \mathcal{T} . A binary temporal interval relation R is defined as $R \subseteq \mathcal{I} \times \mathcal{I}$. In the following we use the notation IRJ for $(I, J) \in R$.

Allen defined a set of seven basic interval relations on two time intervals. Together with the inverses of these relations we obtain a minimal set of relations that can express any qualitative relation between two intervals. This set of relations with respect to the fixed set of intervals \mathcal{I} will be denoted \mathcal{B} and the thirteen relations therein are $\mathcal{B} = \{b, \bar{b}, m, \bar{m}, o, \bar{o}, s, \bar{s}, d, \bar{d}, f, \bar{f}, eq\}$, where \bar{r} is the inverse relation of r .

Definition 3. Let \mathcal{B} be the set of basic relations on any two intervals $I, J \in \mathcal{I}$, with I^- denoting the start point and I^+ the end point of interval I , and similarly for J :

| | |
|---|-------------------------------------|
| $IbJ \Leftrightarrow (I^+ < J^-)$ | Interval I is before interval J |
| $ImJ \Leftrightarrow (I^+ = J^-)$ | Interval I meets J |
| $IoJ \Leftrightarrow (I^- < J^-) \wedge (I^+ < J^+) \wedge (J^- < I^+)$ | Interval I overlaps J |
| $IsJ \Leftrightarrow (I^- = J^-) \wedge (I^+ < J^+)$ | Interval I starts J |
| $IdJ \Leftrightarrow (J^- < I^-) \wedge (I^+ < J^+)$ | Interval I is during J |
| $IfJ \Leftrightarrow (J^- < I^-) \wedge (I^+ = J^+)$ | Interval I finishes J |
| $IeqJ \Leftrightarrow (I^- = J^-) \wedge (I^+ = J^+)$ | Interval I is equal to J |

The inverse of a relation R is denoted \bar{R} , and is defined as $I\bar{R}J \equiv JRI$, the relation equals is thus its own inverse.

In the examples we use events with an interval index instead of pure interval expressions, as this simplifies the exposition. The formal details are deferred until Section 3.

Example 1. Consider our lung disease example. A certain group of COPD patients tends to have relatively frequent exacerbations – events of worsening of symptoms – that are usually caused by airway infections. Using the basic temporal relations we can describe that an infection in interval Inf_I at least partially precedes the increase in symptoms in interval Sym_J , where the notation indicates the event and the interval in which it occurs. We then obtain the statement: $\text{Inf}_I o \text{Sym}_J$, which means that symptoms can outlast the infection. Since an exacerbation is defined as an increase of the relevant symptoms in the interval we can say: $\text{Exa}_K eq \text{Sym}_J$.

Definition 4. An Allen relation is defined as a disjunction of basic interval relations, represented as a set. The power set of the basic relations (all Allen relations) is denoted $\mathcal{A} = \wp(\mathcal{B})$. An interval formula is then of the form IRJ with I, J intervals and $R \in \mathcal{A}$.

Because we will be using Allen's relations as logical relations in what follows, it is useful to notice the effects of Boolean operations on basic relations. The definition above states that Allen's relations are disjunctions of basic relations. By the completeness of the basic relations we have that conjunctions of basic relations are false by definition (at most one relation can hold between any two intervals). For the negation of a basic relation $R \in \mathcal{B}$ we obtain $\neg R = \mathcal{B} \setminus \{R\}$. Note that the negation is thus different from the inverse \bar{R} .

Example 2. COPD patients often have what is called ventilation-perfusion inequality – a mismatch between air flow and blood flow through the lung – which may develop during an exacerbation due to increased airway obstruction. When an exacerbation occurs we have an interval Vpi_I which is during, finishes or is overlapped by Exa_J . Without any further information the relation between ventilation-perfusion inequality and exacerbation can thus be described by: $\text{Vpi}_I\{\bar{o}, d, f\}\text{Exa}_J$.

2.2 Logical reasoning with the interval algebra

As Allen showed [7], this qualitative algebra is well suited to reason about time in a logic context. We are thus abstracting somewhat from the relational perspective above, and proceed to use a logical framework, that is, Allen’s relations are represented by temporal predicates. The logic we will be using derives from the logic programming tradition of using Horn clauses, $H \leftarrow B_1, \dots, B_n$, where H is the head of the clause and B_1, \dots, B_n the body and H and the B_i are logical atoms. Variables are denoted with upper case and are implicitly universally quantified, conjunctions are denoted by commas ‘,’ and a semicolon ‘;’ denotes a disjunction, as in Prolog.

Also instead of using a reified logic approach as Allen does (i.e. using meta-predicates like `HOLDS`), we opt for the arguably simpler framework of temporal arguments [8]. This means that temporal predicates have a temporal argument specifying the relevant time interval. Note that this implies a typed logic, which we will leave implicit as this can always be translated to first order logic at the cost of notational convenience.

Example 3. Consider again our COPD example, now in logic representation:

$\text{exacerbation}(P, I') \leftarrow \text{patient}(P), \text{infection}(P, I), o(I, I').$
 $\text{vpi}(P, I') \leftarrow \text{patient}(P), \text{exacerbation}(P, I), (\bar{o}(I', I); d(I', I); f(I', I)).$

Here `vpi` stands for ventilation-perfusion inequality.

2.3 CP-logic

To represent and reason with probabilistic knowledge, we will use the probabilistic logic language CP-logic [6]. This language is based on Prolog – providing the logic part of the language – extended with probabilistic semantics. The main intuition is that probabilistic logic statements represent causal laws, that is a logic clause gives a relation from some cause to a set of possible outcomes (each with some probability).

Definition 5. A causal probabilistic law has the form: $(H_1 : \alpha_1) ; \dots ; (H_n : \alpha_n) \leftarrow B$ where α_i is the (non-zero) probability of outcome H_i such that $\sum_{i=1}^n \alpha_i \leq 1$; H_i are logical atoms and B is the body of the clause.

In other words, a causal law gives a distribution over possible effects of a cause B . CP-logic is restricted to finite domains, so although you can write quantified rules, these are expanded to a set of ground instances for reasoning. The probabilistic semantics of CP-logic can be described as follows. As is common in logic programming the semantics are defined in terms of Herbrand interpretations, that is the domain is the set of constants of the theory and a symbol is interpreted as itself. The Herbrand universe is the set of all ground terms and the Herbrand base the set of ground atoms.

Definition 6. Let H_U denote the Herbrand universe. A probabilistic process over H_U is a pair $\langle \mathbf{T}, \mathbf{I} \rangle$, where $\mathbf{T} = \langle V, E \rangle$ is a tree with each edge $e \in E$ labelled with a probability $P((v, w))$ and where for each node $v \in V$ the probabilities of the outgoing edges of v sum to 1: $\sum_{w:(v,w) \in E} P((v, w)) = 1$; \mathbf{I} is an interpretation function, mapping nodes in \mathbf{T} to Herbrand interpretations, i.e. subsets of the Herbrand base.

Each transition between nodes is a probabilistic event, described by a causal law.

Definition 7. A causal law c fires in a node v of \mathbf{T} if v has child nodes v_1, \dots, v_{n+1} such that for $1 \leq i \leq n$: $\mathbf{I}(v_i) = \mathbf{I}(v) \cup \{H_i\}$ and the label of the edge (v, v_i) is α_i ; $\mathbf{I}(v_{n+1}) = \mathbf{I}(v)$ and the label of the edge (v, v_{n+1}) is $1 - \sum_i \alpha_i$.

The leaves of a probability tree each describe a possible outcome of the events modelled by causal laws. The probability of a leaf node l is the product of the labels on the edges from l to the root of the tree. By the fact that each causal law fires independently, the product over the probabilities of the outcomes on the path is indeed the probability of the state in l given the events on the path. Since there may be multiple series of events that lead to the same final state the probability of an interpretation is the sum over all the leaves in the tree that share the same interpretation. See also Vennekens et al. [6].

Example 4. In Figure 2 a CP-logic event tree is shown, representing the situation of whether a COPD patient suffers an exacerbation caused by either an infection or by breathing in a noxious substance. The tree follows from these CP-laws:

exacerbation : 0.6 \leftarrow infection.
 exacerbation : 0.2 \leftarrow noxious_substance.
 infection : 0.05.
 noxious_substance : 0.01.

The probability of an exacerbation can be computed by summing over the leaves $l \in V$ that contain E (short for exacerbation) on the path from the root to l . The probability of, for instance, the left most path is $0.05 \cdot 0.6 \cdot 0.01 \cdot 0.2 = 0.00006$ and the probability of an exacerbation is:

$$0.00006 + 0.00024 + 0.0297 + 0.00004 + 0.0019 = 0.03194.$$

Another way to obtain the probabilities is by considering a probability distribution over ground logic programs, which is the usual interpretation of ProbLog, a probabilistic language related to CP-logic, see e.g. [5]. Each grounding of a fact c_i in the logic theory T has some probability p_i , thus given that we only consider finite theories, a finite number of substitutions θ_{ij} gives the grounded set of all logical facts from the theory $L_T = \{c_1\theta_{11}, \dots, c_1\theta_{1j_1}, \dots, c_n\theta_{n1}, \dots, c_n\theta_{nj_n}\}$. The program defines a probability distribution over ground logic programs $L \subseteq L_T$:

$$P(L \mid T) = \prod_{c_i\theta_j \in L} p_i \prod_{c_i\theta_j \in L_T \setminus L} (1 - p_i) \quad (1)$$

This is equivalent with the product over edges in the event tree described above. The probability of a query q is the marginal of $P(L \mid T)$ with respect to q , that is, those

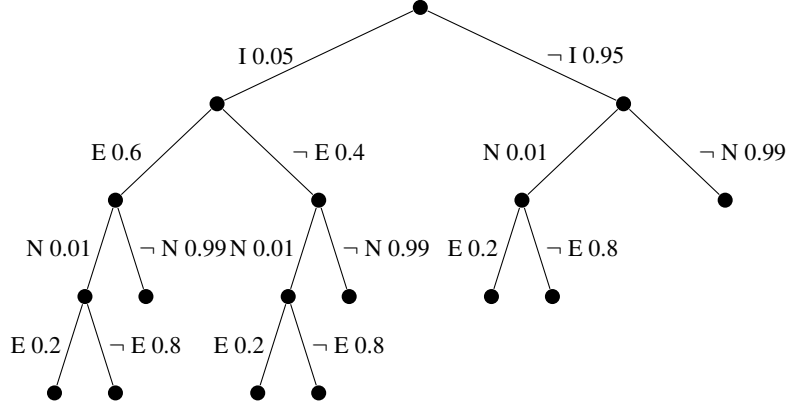


Fig. 2. A probability tree, where I is short for the *infection* event, N denotes *noxious substance* and E is *exacerbation*.

groundings of the program that prove the query $BK \cup L \models q$, where BK is the background knowledge. If we take $P(q \mid L) = 1$ if there exists a substitution θ such that $BK \cup L \models q\theta$ and $P(q \mid L) = 0$ otherwise, we obtain:

$$P(q \mid T) = \sum_{L \subseteq L_T} P(q \mid L)P(L \mid T), \quad (2)$$

which is again equivalent with the probability tree view, where we sum over the leaves in the tree that have the same interpretation.

3 A probabilistic extension of Allen's algebra

Probabilistic logic is a useful tool that combines logical and probabilistic reasoning, which can be used to extend Allen's logic framework for qualitative temporal reasoning with uncertainty. When modelling real world situations, qualitative time is useful for those processes for which it is difficult to obtain precise timing information. However, even when available timing information is only qualitative, events that occur are not necessarily equally likely. It may be possible to obtain likelihood information, telling us that some event is more likely to happen at a particular time, even when the timing information is imprecise. This kind of information can be represented using probabilistic logic. Temporal process descriptions – as represented with Allen's logic – can therefore be extended with probabilistic information, and the capabilities of CP-logic will appear sufficient to act as a basis for such an extended, qualitative temporal and uncertain logic.

3.1 On events and intervals

To model uncertain processes we are primarily interested in the occurrence of events. In our context we consider events that are uniquely associated with intervals, and this unique association is expressed by means of a time-interval index.

Definition 8. Let \mathcal{E} denote the event space containing all probabilistic events of interest. A temporal event E_I is defined as a probabilistic event $E \in \mathcal{E}$ that is temporally uncertain, expressed by the time interval $I \in \mathcal{I}$.

In the probabilistic logic context sketched in the previous section events are represented by facts, which are interpreted in probabilistic logic as independent events. Relations between facts are stated by logical expressions and through these expressions, dependences between events embedded in the facts can be introduced. Now to define temporal events, we index facts with time, which in our case means that facts are augmented with an interval valued argument.

An important issue is the interpretation of events associated with intervals. At least three interpretations seem possible:

- An event implies that somewhere during the interval the modelled occurrence happens. That is the event is instantaneous but not precisely defined in time.
- An event is some occurrence that has a certain duration, for which the exact temporal extend is unknown. The interval gives the temporal bounds in which the event is contained.
- An event occurrence lasts the complete duration of the interval.

These interpretations lie on a spectrum from instantaneous events to extended events with different levels of temporal uncertainty involved. For instantaneous events uncertainty about a time interval can be derived from the uncertainty of the time points. However, in many domains, medicine among them, it is unrealistic to model events as instantaneous. When considering events with duration for which the interval gives lower and upper bounds for the start and end of the event, there can be uncertainty about the interval bounds and uncertainty about when the event occurs within the interval. For events that have a duration equal to the interval length, the temporal uncertainty lies in which interval is assigned to the event, since there is no uncertainty inside the interval.

As it is possible to have intermittent events, that is a recurring process which could be seen as a single event, we need to define how events with multiple intervals interact. For example, a fever may abate for a day and then return, which one could still look on as a single fever event. The algebraic properties of temporal events should make clear which properties hold. The Boolean algebra of temporal events $B(\mathcal{E}_{\mathcal{I}})$, where $\mathcal{E}_{\mathcal{I}}$ is defined as $\mathcal{E}_{\mathcal{I}} = \{E_I \mid E \in \mathcal{E}, I \in \mathcal{I}\}$, should obey certain rules, taking into account the time interval indices of the events. The elements of the Boolean algebra are obtained by constructing conjunctions of events $(E_I \wedge E'_J)$, disjunctions $(E_I \vee E'_J)$ and negations $\neg E_I$, with events $E_I, E'_J \in \mathcal{E}_{\mathcal{I}}$.

Now, for $E = E'$ with temporal events E_I and E'_J , there is an interaction between the Boolean operations on events and the relation between the time intervals. For example, if $IeqJ$ holds, then $(E_I \wedge E_J) = E_I = E_J$. In addition, when two intervals I and J meet, i.e. ImJ holds, then $(E_I \wedge E_J) = E_K$ with $K = I \cup J$. Thus the event E actually occurred during interval K . It turns out that $(E_I \wedge E_J) = E_{I \cup J}$ holds for $I * J$ with $*$ $\in \mathcal{B} \setminus \{b, \bar{b}\}$, i.e. it holds for all cases except when I and J are disconnected intervals. The other operations of the Boolean algebra do not interact with the relations.

3.2 Uncertainty

Uncertainty can be incorporated into the resulting language with respect to the temporal events E_I ; and with regard to the relation between time intervals IRJ . When combined for two temporal events E_I and E'_J , these assumptions would give rise to a joint probability distribution of the form $P(E_I, E'_J, IRJ)$ with R an Allen relation. Note the special case when a single event E is associated with both intervals.

By conditioning on IRJ , one removes part of the uncertainty, yielding: $P(E_I, E'_J \mid IRJ)$. We first look at this case where only events are uncertain, and relations between interval are considered to be part of the logic specification. Because events are associated with intervals, the logic relation serves as a constraint on the events and as such still influences the uncertain part. Let us now look at specifying probabilities for events.

Definition 9. *The probability of a Boolean expression of events is given by the probability function $P : B(\mathcal{E}_{\mathcal{T}}) \rightarrow [0, 1]$, where $B(\mathcal{E}_{\mathcal{T}})$ denotes the Boolean algebra over the set of temporal events $\mathcal{E}_{\mathcal{T}}$.*

This defines uncertainty in terms of what is sometimes called a probability algebra – a pair consisting of a Boolean algebra and a probability function. But it tells us nothing yet over the actual shape of the distribution; one could for example parametrise on interval properties such as start point and length, or some domain dependent parametrisation.

We may also be interested in answering questions about whether a certain event occurred irrespective of time. For this we need a concept of atemporal events, summarising over the intervals. This is also related to multiple granularities, where some events occur at a different time scale than others, which appears to be an intermediate level of summary over time. If we are interested in the probability that an event occurs irrespective of time, we can say that this is equivalent to the probability that the event occurs in at least one interval, which can be expressed by the disjunction over events for each possible interval. Or more formally: $P'(E) = P(\bigvee_{I \in \mathcal{I}} E_I)$, where P' denotes the distribution over atemporal events. Since the disjunction over events is a proper probability, we obtain a distribution over atemporal events.

A similar question which may also be of interest is the probability of whether an event will have occurred by a certain time $t \in \mathcal{T}$. This can be solved by considering a subset of all intervals that satisfy the time constraint. Hence, $P''(E_{<t}) = P(\bigvee_{I \in S} E_I)$ with $S = \{I \mid I \in \mathcal{I}, I < t\}$, where at least all intervals that are entirely before t should be taken into account, hence $I^+ < t$. Limiting S to these intervals gives a lower bound. The meaning of $I < t$ for intervals with $I^- < t < I^+$ depends on the interpretation one chooses for temporal events (see Section 3.1). For the interpretation of events during the whole interval, it depends on whether you are interested in events that have started but not yet finished or only in completed events. For the other two interpretations it may be possible to compute a probability of partial intervals by looking inside the intervals.

Specific distributions If we now choose to define a distribution based on the endpoints of the interval associated with an event E , we can write: $P(E_I) = P_{I^-, I^+}(E_I)$, which indicates that the distribution depends on the parameters I^- and I^+ , which we could also parametrise as $P(E_I) = P_{s, l}(E_I)$, with $s \in \mathcal{T}$ the interval start point and $l =$

$I^+ - I^-$ the length of the interval. Hence, the probability of for example an exacerbation event depends on when the exacerbation starts and its duration, which appear to be reasonable parameters to model clinical events of interest.

Given the parametrisation, we still have to choose the exact shape of the distribution, which depends on the exact situation we want to model. Here we consider a fairly general case, that *when* something happens is often unrelated to its *duration*, resulting in the assumption that the start point and duration are chosen independently, according to their own distribution.

Definition 10. *Let P be a specific distribution following Definition 9. Assuming that the probability of an event is described by the probability of the start point and duration of the interval, we obtain $P(E_I) = P_s(E_I)P_l(E_I)$, with s the interval start point and l the interval length.*

We can now choose a specific distribution for interval start points and interval length. A distribution like, for example, the beta-distribution seems useful to model particular situations as it produces different shapes depending on the parameters. In the discrete case with bounded support we obtain the same flexibility by using a beta-binomial distribution, a combination of a beta-distribution:

$$f(p; \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \quad \text{with} \quad B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx,$$

where $f(p; \alpha, \beta)$ is parametrised by α, β and $B(\alpha, \beta)$ is the beta-function; and the binomial with the well-known form: $f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$. When compounded the beta-distribution gives the probability of the parameter p of the binomial, which leads to:

$$f(k; n, \alpha, \beta) = \binom{n}{k} \frac{\int_0^1 p^{k+\alpha-1} (1-p)^{n-k+\beta-1} dp}{B(\alpha, \beta)} = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)}.$$

The advantage of this distribution is that by choosing the parameters appropriately, we obtain useful special cases like a Bernoulli distribution when $n = 1$; the discrete uniform distribution for $\alpha = \beta = 1$; and a binomial distribution for large α and β .

For any particular modelling situation other distribution may be appropriate. For instance a combination of an exponential distribution (geometric in the discrete case) for interval start and a normal distribution over durations may be useful. But in general the choice of distribution is largely domain specific.

3.3 Reasoning with probabilistic intervals

Reasoning with temporal relations can now be given the added dimension of reasoning with uncertainty by considering the probabilities over temporal events. The logic framework allows us to do temporal reasoning which due to the probabilistic semantics automatically also gives us the probabilities, for which we only need specify probability distributions over events as described above. In medicine a fairly natural question to ask is what the probability of some event is given some other event that is temporally

related, which we can write as the probability: $P(E'_J, E_I : IRJ)$, because IRJ is determined logically it should be interpreted as a constraint on the distribution rather than probabilistic conditioning, hence the notation with a colon. The following proposition shows how we can compute this probability via the CP-logic semantics, given that IRJ is a logical relation between the unknown intervals associated with the events.

Proposition 1. *Let R be a relation $R \in \mathcal{A}$, and $E, E' \in \mathcal{E}$ events; the probability $P(E, E' : IRJ)$ is then obtained as follows:*

$$P'(E, E' : IRJ) = \sum_{I, J} \mathbb{I}(IRJ) P(E_I, E'_J) = \sum_{\ell} \mathbb{I}_{\ell}(IRJ) \prod_{e_{\ell}} P(e_{\ell})$$

where ℓ denotes a leaf node in the tree, e_{ℓ} is an edge on the path from ℓ to the root of the tree; and $\mathbb{I}_{\ell}(x)$ is an indicator function that is 1 if x is true in ℓ .

Proof. The indicator function ensures that R holds. By the fact that if IRJ holds in some leaf node of the tree, the temporal events E_I, E'_J must have occurred on the path from the root to ℓ , hence the probability of the events follows from the probability tree semantics of CP-logic. This can be seen by considering the probabilistic process given in Definition 6 and the transitions between nodes of Definition 7.

Although this proposition shows the probability calculus in the tree representation, it is also informative to see how they follow from the logical reasoning. We can then restate the proposition as follows.

Proposition 2. *Let R be a temporal relation $R \in \mathcal{A}$ and C the set of clauses $\{E(I), E'(J), IRJ\}$, the probability $P(C)$ then follows from the proofs $BK \cup L \models C$.*

Proof. By Equation 2 we sum over all proofs of C , that is all ground substitutions θ such that $BK \cup L \models C\theta$. The probability of each proof is given by Equation 1. Since each proof of IRJ requires instantiations for I and J , there will be probabilistic events E_I, E'_J that are part of the proof and that adhere to the constraint IRJ . Hence we obtain Proposition 1.

Example 5. We are again interested in modelling the relation between an infection and the occurrence of an exacerbation. The temporal relation is still *overlaps*, but now we also have probabilistic information attached to intervals. By choosing the parameters of the beta-binomial distribution $\alpha = \beta = 1$ we obtain a uniform distribution, which we can make explicit by writing down the possible intervals \mathcal{I} with the time-line for this example restricted to $[0, 3)$. We then obtain the following logic specification:

```

o(I1, I2) ← i(I1, S1, E1), i(I2, S2, E2), S2 > S1, S2 < E1, E2 > E1.
i(i1, 0, 1) : P; i(i1, 0, 2) : P; i(i1, 0, 3) : P;
  i(i1, 1, 2) : P; i(i1, 1, 3) : P; i(i1, 2, 3) : P ← betabinom(P).
i(i2, 0, 1) : Q; i(i2, 0, 2) : Q; i(i2, 0, 3) : Q;
  i(i2, 1, 2) : Q; i(i2, 1, 3) : Q; i(i2, 2, 3) : Q ← betabinom(Q).
exacerbation(i2) ← infection(i1), o(i1, i2).
infection(i1).

```

For the betabinomial distribution with $\alpha = \beta = 1$ we find $P = Q = 0.111$. The astute reader then notices that the probabilities in the example do not sum to 1 as would be expected. The reason for this is a censoring effect that results from our definition of probabilities on intervals. That is, each point on the time line is the start point of an interval with a certain probability. Since an interval cannot end before it starts, the possible end points are limited to those points that follow the start point but precede the end of the time line. Although it would be possible to specify a distribution over those points, that would result in the strange situation that shorter intervals become more likely towards the end of the time line. A more natural solution is thus to consider the end of the time line as a boundary that we cannot look beyond, but which does not limit the possibility of event occurring after the boundary. This results in a truncated distribution where the probability mass that falls beyond the time line is simply discarded, hence leading to a sum over interval probabilities lower than 1.

With this representation we can now answer probabilistic queries about our temporal concepts. The reasoning mechanics of CP-logic will take care of the probabilistic part of the queries. The probability of observing for instance an exacerbation in the interval $[1, 3)$ follows from the probability of an infection in some interval that overlaps with $[1, 3)$, which in this case only leaves the interval $[0, 2)$. The probability tree that is constructed thus contains a single path consisting of the events infection and exacerbation, with the uncertainty modelled through the probabilistic choices for the intervals, i.e.: $P([0, 2)) = 0.111$ and $P([1, 3)) = 0.111$ by the uniform distribution. The final probability is the product over the probabilities of the events in the tree, hence $P(\text{exacerbation}([1, 3))) = 0.111 \cdot 0.111 \cdot 1 \cdot 1 = 0.0123$.

An advantage of this representation is that it is possible to start with a logical expression and add probabilities by defining a distribution over intervals. Besides the temporal probabilistic information, the probabilistic logic framework can also be used to incorporate more general probabilistic facts. For instance, in the example above, the infection predicate can easily be assigned a prior probability, for example $\text{infection}(i1) : 0.1$. This models the situation that the probability of contracting an infection is 0.1, and the timing is distributed according to distribution $i1$.

Example 6. Consider again our running example. Now say we observed an exacerbation in the interval $[1, 3)$. Given the observation we can ask the question what the probability $P(\text{vpi}(i1))$ is given the evidence $i(i2, 1, 3)$. The answer follows from the probability tree where the possible outcomes of $i2$ are replaced by the determined evidence $i(i2, 1, 3) : 1$. Now the reasoning mechanism can simply be applied leading to

$$\begin{aligned} P(\text{vpi}(i1)) &= P(i1)P(i2)P(d(i1, i2) \vee f(i1, i2) \vee \bar{o}(i1, i2))P(\text{vpi}(i1)) \\ &= P(i1)P([1, 3))P(d(i1, [1, 3)) \vee f(i1, [1, 3)) \vee \bar{o}(i1, [1, 3))) \cdot 1 \\ &= P([2, 3))P([1, 3))P(f([2, 3), [1, 3))) = 0.111 \cdot 0.111 \cdot 1 = 0.0123 \end{aligned}$$

Note that in these examples we assume that the specified probability distributions are valid given that some temporal relation holds, which means that we modelled the relation between infection and exacerbation within the context of overlapping time intervals. This works for some situations, but it would also be interesting to look at temporal relation as influencing a distribution, instead of as a constraint. We could for example study how, given an event E_I , the additional information IbJ changes the distribution

of E_J . It is unlikely that we can say in general what the effect of temporal information will be, as this will be domain and event specific, however some regularity is expected.

Let us look at a specific case where we have events E_I, E'_J with IbJ . We could have a number of possible situations, for example, E and E' could be ‘either-or’ events which means that the added information of IbJ results in the probability of E'_J becoming zero because E_I already occurred. In our running example this could be the case for the probability of an exacerbation after the infection has ended (if we leave other causes like a second infection out of consideration). Another situation could be that E_I facilitates the occurrence of E' , thus increasing the probability of E'_J when IbJ holds. Yet another possibility would be two events that usually occur overlapping in time, for which the additional information IbJ makes E'_J less likely.

The pattern that emerges shows some resemblance to the qualitative influences in qualitative probabilistic networks [9], where the temporal relation IbJ has a positive or negative influence on the probability of events given the relation. That is the probability of E'_J increases (decreases) when E_I has a positive (negative) influence given that the relation holds. The problem with such a characterisation is that it is hard to imagine what we can do with this in practice. Knowing that the probability increases does not tell us how exactly we should change the probability distribution. Nevertheless, studying this kind of patterns might be useful from a knowledge representation viewpoint; by defining specific patterns of temporal influence on distributions we acquire an additional modelling tool. So although we cannot ascertain the effect of temporal relations in general, it may be useful to add specific cases to our modelling language.

Uncertainty on relations As mentioned in Section 3.2 we may also be interested in the situation where not only events are uncertain but also the temporal relations themselves. The uncertainty on relations conveys that it may not be known what the order of events is. Allen’s algebra uses disjunctions to model this situation, but with probabilities we can use additional likelihood information, if available, and otherwise a uniform distribution can be used to regain the deterministic case. The joined probability that is now relevant is: $P(E_I, E_J, IRJ) = P(E_I, E_J \mid IRJ)P(IRJ)$. Compared to the previous situation we thus need probabilities defined on relations, for which a discrete distribution constructed by hand seems adequate as the number of probabilities that has to be specified is at most 13 (the cardinality of \mathcal{B}). Often a few relations will be more likely than others limiting the number of probabilities further. Note that although the set \mathcal{A} is large ($|\mathcal{B}| = 2^{13}$) the probability of a relation $R \in \mathcal{A}$ is the sum over the basic relations in R , due to the mutual exclusivity of the basic relations.

Reasoning with uncertain relations requires little additional machinery. It results in additional probabilistic facts in our logic, but these behave like any other fact. The combined logic and probabilistic reasoning thus proceeds as described before. How a prior probability on relations should be interpreted, independent of events, is less clear however. One solution would be using domain knowledge to count how many times certain relations occur when marginalising over all events of interest. This requires a relevant data set, that then provides prior probabilities for the occurrence of temporal relations that are meaningful within the specific domain.

4 Related work

Allen's algebra has had much attention over the years and finds applications in various fields, from planning to clinical medicine and many more. Besides Allen's work [1, 7], well-known contemporary work is the temporal logic by McDermott [10]. Both authors deal with much broader concepts than those considered here, like continuous change, actions and plans, but interesting to note is McDermott's observation that quite a few problems result as a consequence of uncertainty, and that no formal framework exists that satisfactorily combines logic and probability. Fortunately this has changed in recent years, leading to our current work on temporal reasoning in probabilistic logic.

Probabilistic temporal interval networks [11] are a probabilistic extension of the network representation often used to specify the consistency problem in Allen's algebra. The relations between two intervals are weighted with probabilities. This is thus a generalisation of the uncertainty that can exist about what relation holds between two intervals from disjunctions to a distributions over relations. Probabilistic temporal networks [12], are network models that incorporate Allen's constraints on conditional probabilities. They make the assumption that the intervals of interest are known beforehand and can be specified explicitly, thus not allowing uncertainty in what intervals are or will be of interest. Then, recently, probabilistic logic has been applied to represent stochastic processes [13]. The proposed language CPT-L extends CP-logic to represent fully-observable homogeneous Markov processes, and allows efficient reasoning.

5 Conclusion

In this paper we have looked at the interaction of qualitative time and probability, to obtain a temporal representation with uncertainty. The representation language seems usable to reason with uncertain temporal information, although work remains to further study its properties and applicability.

References

1. Allen, J.: Maintaining knowledge about temporal intervals. *Commun ACM* (1983)
2. Shahar, Y.: A framework for knowledge-based temporal abstraction. *Artif Intell* (1997)
3. Poole, D.: Probabilistic Horn abduction and Bayesian networks. *Artif Intell* (1993)
4. Kersting, K., de Raedt, L.: Bayesian logic programs. TR 151, University of Freiburg (2000)
5. Kimmig, A., Demoen, B., de Raedt, L., Santos Costa, V., Rocha, R.: On the implementation of the probabilistic logic programming language ProbLog. *Theor Pract Log Prog* (2010)
6. Vennekens, J., Denecker, M., Bruynooghe, M.: CP-logic: A language of causal probabilistic events and its relation to logic programming. *Theor Pract Log Prog* (2009)
7. Allen, J.: Towards a general theory of action and time. *Artif Intell* (1984)
8. Bacchus, F., Tenenber, J., Koomen, J.: A non-reified temporal logic. *Artif Intell* (1991)
9. Wellman, M.: Fundamental concepts of qualitative probabilistic networks. *Artif Intell* (1990)
10. McDermott, D.: A temporal logic for reasoning about processes and plans. *Cog Sc* (1982)
11. Ryabov, V., Trudel, A.: Probabilistic temporal interval networks. In: *TIME*. (2004)
12. Santos Jr., E., Young, J.: Probabilistic temporal networks: A unified framework for reasoning with time and uncertainty. *Int J Approx Reason* (1999)
13. Thon, I., Landwehr, N., de Raedt, L.: Stochastic relational processes: Efficient inference and applications. *Mach Learn* (2011)

Effective priors over model structures applied to DNA binding assay data

Netherlands Cancer Institute, Amsterdam
Nicos Angelopoulos and Lodewyk Wessels

Abstract. This paper discusses Distributional Logic Programming (Dlp), a formalism for combining logic programming and probabilistic reasoning. In particular, we focus on representing prior knowledge for Bayesian reasoning applications. We explore the representational power of the formalism for defining priors of model spaces and delve to some detail on generative priors over graphical models. We present an alternative graphical model learnt from published data. The model presented here is compared to a graphical model learnt previously from the same data.

1 Introduction

The ability to represent complex prior knowledge would greatly benefit the application of Bayesian methods as it can focus computational resources in areas that of particular interest as expressed by the prior. Furthermore, Bayesian methods provide a convenient, clean framework in which such knowledge can be incorporated. Expressing complex prior knowledge and its incorporation within Bayesian statistics is thus an important and promising line of research.

Logic programming (LP) is an attractive formalism for representing crisp knowledge. Its basis on formal mathematical logic lends it strong affinity to the very long tradition in research in knowledge representation and logical reasoning. However, the boolean nature of first-order logic sits uncomfortably with modern approaches to epistemic inference, which are statistical in nature. To remedy this, a number of probabilistic extensions to LP have been proposed. Of particular interest to this contribution are those that have been introduced for the purpose of representing Bayesian priors ([5, 1]). Here, we present an extension to the probabilistic aspects of their formalism based on probabilistic guards which have been used in more abstract probabilistic languages such as PCCP ([7]).

Currently, biology is an area of scientific knowledge that is expanding at an unprecedented rate. Vast volumes of data is being generated and knowledge in the form of scientific papers is being accumulated. Invariably, however, statistical analysis is performed *ad hoc* and knowledge is considered only implicitly in the form of assumptions. These can not be precise or quantitative. By incorporating existing knowledge in a disciplined framework computational and statistical inference can be guided to the areas that are still lacking evidence and drive the construction of more precise models. Knowledge based data analysis is [12]

We demonstrate the usefulness of our formalism by applying MCMC inference over graphical models on a published dataset. For comparison, we use a fairly agnostic prior. The MCMC consensus graph constructed from ten MCMC chains, is in broad agreement with that learnt by the bootstrapping methods described in [8] and implemented in the Banjo software [15]. The bootstrapping graph was taken from the literature [16]. The use of stronger priors would further benefit our approach, where as only simple information, such as absence/presence of specific edges can be incorporated in most other systems.

2 Preliminaries

A logic program L is a set of clauses of the form $Head \text{ :- } Body$ defining a number of predicates. $Head$ is a single positive literal or atom, constructed from a predicate symbol and a number of term arguments. Each term is a recursively defined structure that might be an atomic value, a variable or a function constructed by an atomic function symbol and n term arguments. A query or goal G_i is a conjunction of literals $(A_{(i,1)}, \dots, A_{(i,n)})$ which the logic engine attempts to refute against the clauses in L . This is done by employing SLD with a top to bottom scan of the clauses as the rule. Linear resolution at step i will resolve $A_{(i,1)}$ with the head (H_i) of a matching clause (M_i) and replace it with the body of the clause. The step at $i + 1$ is recursively defined by applying resolution to the newly formed goal. Matching is via the unification algorithm, which when successful, provides a substitution θ_i such that $A_{(i,1)}/\theta_i = H_i$. A computation terminates when the current goal is the empty one or a failure to match any clause occurs. The logic engine can be used to explore unreached parts of the space by returning to the latest matching step and attempting to find alternative resolution clauses. The full search ends when all alternatives have been exhausted. In what follows we will use A_i to refer to $A_{(i,1)}$, i.e. the atom used for the i th resolution step. As an illustrating example of a logic program consider the following two clauses defining the *member/2* relation (also referred to as predicate):

$$\begin{aligned} (C_1) \text{ member}(H, [H|T]). \\ (C_2) \text{ member}(El, [H|T]) \text{ :-} \\ \quad \text{member}(El, T). \end{aligned}$$

The first clause (C_1) states that the head of a list is one of its members, while the second one states that element El is a member of the list, if it is a member of the tail (T) of the list. Lists are convenient recursive structures term structures commonly used in logic programming to hold a collection of terms. Posed with a query of the form $? - \text{member}(X, [a, b, c])$ the LP engine will use SLD resolution which scans the query left to right and the program top to bottom as to provide all possible answers in the form of alternative values for X .

2.1 Probability Theory and Logic Programming

Logic programming implements a systematic search of a non-deterministic space. In this paper we will review some of the difficulties of mixing such spaces with probabilistic ones and present one way to achieve this. The thesis we propose is that any formalism which treats probabilities as top-level constructs, must define a single probabilistic space and in the case of logic programming a clear distribution over Θ . Current approaches to probabilistic formalisms include: the replacement of non-determinism by a probabilistic operator, the use of a primitive that appears within limited non-determinism and a clear separation of the two spaces. First, SLPs [11] under the semantics presented in [5] replace SLD resolution with sampling over pure programs that only contain stochastic clauses. An example of the second category is Prism, [14]. It provides a single probabilistic construct that instantiates an unbound variable from the elements of a list according to the probability values attached to each element. It was introduced with parameter learning in the

context of PCFGs (Probabilistic Context Free Grammars) and hidden Markov models in mind. PCLP [13] and $\text{clp}(\text{pfd}(Y))$ [2] employ constraint programming to, in distinct ways, create two separate spaces. The non-determinism remains within the clausal level while the probabilistic is constructed in the constraint store with the constraint solver used to reason/infer from this information.

3 Syntax

We extend the clausal syntax with probabilistic guards that associate a resolution step to a probability which is computed on-the-fly. The main intuition is that in addition to the logical relation a clause defines over the objects in its head arguments it also defines a probability distribution over aspects of this relation.

Definition 1. *Probabilistic clauses in Dlp are a syntactic extension of definite clauses in LP. Let $Expr$ be an arithmetic expression in which all variables appear in the clause-unique unary functions of the comma separated tuple $GVars$. Let $Guard$ be a goal and $PVars$ be a comma separated tuple of variables that appear in $Head$. A probabilistic clause is defined by:*

$$Expr : GVars \cdot Guard \sim PVars : Head :- Body \quad (1)$$

Arithmetic expressions of clauses defined by (1) will be evaluated at resolution time. In cases where this can be done successfully, the clauses will be used to define a distribution over the probabilistic variables ($PVars$). The distribution may depend on an arbitrary number of input terms via calls to the guard.

We also allow goals that appear in the body of clause definitions to be labelled by a tuple of unary functions each wrapping an arithmetic expression. Each of the unary functions corresponds to the functions in $GVars$. The intuition behind labelled goals in the body of clauses ($Body$) is that often probability labels of recursive calls can be easily computed from their parent call thus the interpreter can avoid recomputing all or some of the guards. For a single probabilistic predicate all clauses must define the same set of probabilistic variables. In what follows we let C_i^\sim denote the set of probabilistic variables of clause C_i . Comparing to the already introduced $member/2$ relation, the following is a probabilistic version $pmember/2$.

$$\begin{aligned} (C_3) \quad & \frac{1}{L} : l(L) \cdot \text{length}([H|T], L), 0 < L \sim H: \\ & \text{pmember}(H, [H|T]). \\ (C_4) \quad & 1 - \frac{1}{L} : l(L) \cdot \text{length}([H|T], L), 0 < L \sim El: \\ & \text{pmember}(El, [H|T]) :- l(L-1): \text{pmember}(El, T). \end{aligned}$$

These clauses have attached to them expressions which will be computed at resolution time. (C_3) is labelled by $\frac{1}{L}$ where L is the length of the input list (as defined by standard predicate $list/2$ which is present in all prolog systems). (C_4) claims the residual probability. The recursive call has been augmented to carry forward the value of L as the length of T is one less than that of the input list and thus we avoid recomputing the guard. Intuitively, for the query $? - pmember(X, List)$ where $List$ is a known list, the program defines an equiprobable distribution over all the possible element selections from the list. The three corresponding probabilities when $List = [a, b, c]$ are computed as $\frac{1}{3}, \frac{2}{3} \times \frac{1}{2}, \frac{1}{3} \times \frac{1}{2} \times 1$. Clauses

(C_3) and (C_4) are written in full syntax. It is often unnecessary to be as verbose. We will drop the unary function from variables that are named with the upper case version of the functor name, that is in our example $l(L)$ reduces to L . We will share guards among clauses, thus the guard part of (C_4) can be removed. We also drop the unary functor from guard variables in body calls when either the corresponding predicate has a single guarded variable or a single guarded input variable is involved in an expression. Finally, in the interest of clarity we introduce guard lines to our programs which factor the guard section out. The example program is then:

$$\begin{aligned}
(G_1) \quad & L \cdot \text{length}(\text{List}, L), 0 < L \sim \text{El} : \text{pmember}(\text{El}, \text{List}). \\
(C'_3) \quad & \frac{1}{L} : G_1 : \text{pmember}(H, [H|T]). \\
(C'_4) \quad & 1 - \frac{1}{L} : \quad \text{pmember}(\text{El}, [H|T]) :- \\
& \quad \quad \quad L-1: \text{pmember}(\text{El}, T).
\end{aligned}$$

A distributional logic program R is the union of a set of definite clauses L and a set of distributional clauses D , defining the logical and probabilistic parts of the program respectively. D and L must define disjoint sets of predicates. Dlp grew out of the need to extend SLPs. Having labels that are set numbers has the major advantage that parameter learning can be done efficiently, [6], but can not describe complex probabilistic dependencies. For instance, $\text{pmember}/2$, as defined above cannot be capture by a simple stochastic logic program. The fact that SLPs labels are fixed numbers means that they cannot encode the uniform choice of a list element in a linear fashion by traversing the list.

4 Priors over graphical models

Hereafter, we will use the terms graphical model and BN, for Bayes nets, interchangeably. A graphical model can be easily represented as a LP term. For instance, the structure of a BN with nodes 1, 2 and 3, and two parent edges from 1 to 3 and from 2 to 3 corresponds to the term structure $[1 - [3], 2 - [3]]$. (Stochastic) logic programs can be written that define the space of all models, say all possible BNs with N nodes. The benefit of doing so, is that the high-level and theoretically sound properties of logic programs can provide a suitable platform for representing domain knowledge.

One approach to constructing the dependency graph of a BN is by recursively choosing parents for each of the possible nodes. Care must be taken however as to avoid introducing cycles in the graph. This method is well suited to situations where prior information regarding edges in the graph is available. The top level non-stochastic part of the selection expressed in logic programming is :

$$\begin{aligned}
(B_2) \quad & \text{bn}([], _Nds, []). & (B_1) \quad & \text{bn}(Nds, BN) : - \\
(B_3) \quad & \text{bn}([Nd|Nds], AllNds, BN) : - & & \text{bn}(Nds, Nds, BN), \\
& \quad \quad \quad \text{parents_of}(Nd, AllNds, Pa), & & \text{no_cycles}(BN). \\
& \quad \quad \quad BN = [Nd - Pa|BN], \\
& \quad \quad \quad \text{bn}(Nds, AllNds, BN).
\end{aligned}$$

Given a list of possible nodes Nds that appear in BN , predicate $\text{bn}/2$ constructs a candidate graph and then checks that the graph produced includes no cycles. If that is not

the case, the program fails. Predicate $bn/3$ traverses the nodes selecting parents for each one of them from $AllNds$. When an ordering is known over the variables in the BN, its construction can proceed without checking for cycles. The ordering constraint [8] specifies that the order of nodes (Nds) is significant and that each node can only have parents from the section of the ordering that follows it.

$$\begin{aligned}
(B_5) \quad & bn([], _AllNds, []). & (B_4) \quad & bn(Nds, BN) : - \\
(B_6) \quad & bn([Nd|Nds], PossPa, BN) : - & & bn(Nds, [], BN). \\
& \quad \quad \quad parents_of(Nd, PossPa, Pa), \\
& \quad \quad \quad BN = [Nd - Pa|TBN], \\
& \quad \quad \quad bn(Nds, [Nd|PossPa], TBN).
\end{aligned}$$

Clauses $(B_4 - B_6)$ provide a compact implementation for the ordering constraint. The program is also robust in relation to the probabilistic paths associated to the model instances they generate. Each model has a unique non-probabilistic part with regard to this program segment and it never leads to a failure. On the contrary clauses $(B_1 - B_3)$ lead to failure and loss of probability mass when a cycle is introduced. This can only be detected after some probability is assigned to the failed path. It is worth noting that clause (B_6) selects parents for a node from the set of possible parents rather than the set of all nodes. Also, when the ordering is not known (program $B_1 - B_3$) there is no good reason why child variables should be selected in sequential order. The following program puts the two ideas in use:

$$\begin{aligned}
(B_8) \quad & bn([], _All, BN, BN). & (B_7) \quad & bn(Nds, BN) : - \\
(B_9) \quad & bn(Nds, All, BnSoFar, BN) : - & & bn(Nds, Nds, [], BN). \\
& \quad \quad \quad pmember(Nds, Nd, RemNds), \\
& \quad \quad \quad poss_pa(Nd, BnSoFar, All, PossPa), \\
& \quad \quad \quad parents_of(Nd, PossPa, Pa), \\
& \quad \quad \quad add(Nd - Pa, BnSoFar, NextBnSF), \\
& \quad \quad \quad bn(RemNds, All, NextBnSF, BN).
\end{aligned}$$

Here the node is selected probabilistically ($pmember/3$) from the nodes still available. The selection can be either fair or biased. Clause (B_9) uses an auxiliary structure $BnSoFar$ which accumulates the graph of the BN at the current level. This is used by $poss_pa/4$ to eliminate cycle introducing parents. Clause (B_8) terminates the recursion. Once all nodes have been assigned parents, the auxiliary structure is unified to the variable of the complete BN. A number of distributions from the literature can be fitted over the edge selection that connects children in the BN to their parents. [9] introduced $p(BN) \propto \kappa^\delta$ where κ is a user defined parameter and δ is the number of differing edges/arcs between BN and a ‘prior network’ which encapsulates the user’s prior belief about the network structure. [3] suggested a generalisation of the above that allows for arbitrary weights for each missing edge: $p(BN) \propto \sum_{ij} \kappa_{ij}$. A Dlp can encode such information by simply passing a list of length n_i as an extra argument to the BN constructing clauses. Each sublist is a list of length n_j weights that can be used to call the following predicate :

$$\begin{aligned}
(B_{10}) \quad & K_{ij} : parent_edge(PP, [PP|TPa], TPa). \\
(B_{11}) \quad & 1 - K_{ij} : parent_edge(PP, TPa, TPa).
\end{aligned}$$

In the context of learning from expression array data, [17] constructed tabular priors over the existence of some edges. This is complementary to penalising missing edges. A similar program to the one presented above can capture such knowledge.

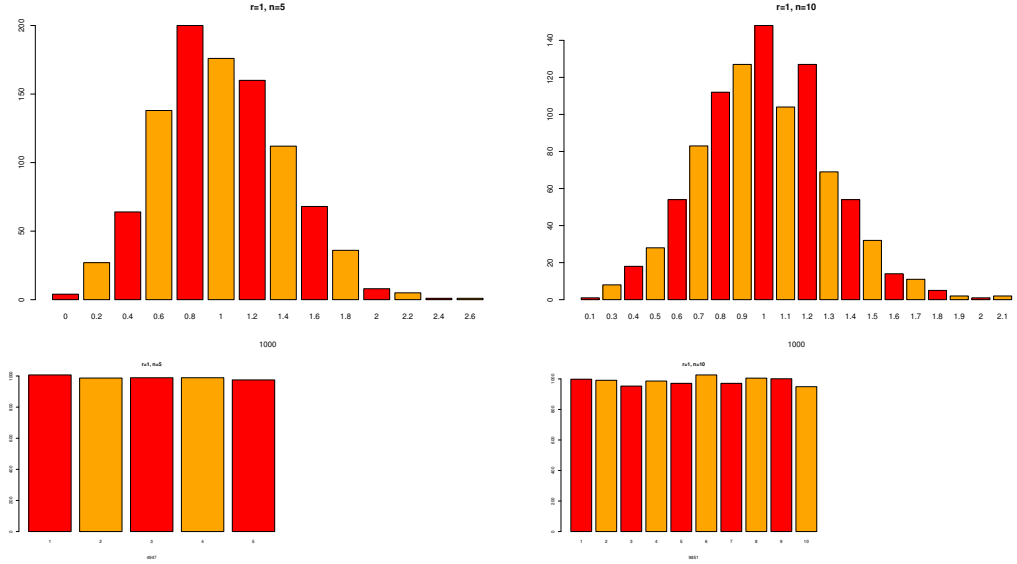


Fig. 1. 1000 samples from clauses B_{12} - B_{17} . Top left: average family size for $r=1, n=5$. Top right: average family size for $r=1, n=10$. Bottom left: number of times node i (x-axis) was a parent, for $r=1, n=5$. Bottom right, as adjacently but for $r=1, n=10$.

Another approach proposed in [8] is that of limiting the number of parents a node may have. It seems natural that a prior distribution over the parental population maybe a suitable extension to this and in Dlp it can be written as :

$$(G_2) \quad \begin{array}{l} L \cdot \quad \text{length}(\text{PossPa}, L1), L \text{ is } L1 - 1 \\ R \cdot \\ \sim Pa : \quad \text{parents_of}(\text{PossPa}, R, Pa). \end{array}$$

$$\begin{array}{ll} (B_{13}) \text{ graph}([], _Nds, _R, G). & (B_{12}) \text{ graph}(Nds, R, G) : - \\ (B_{14}) \text{ graph}([H|T], Nds, R, G) : - & \text{graph}(Nds, Nds, R, G). \\ & \text{select}(Nd, Nds, \text{PossPa}), \\ & \text{parents_of}(Nd, \text{PossPa}, Pa), \\ & G = [H - Pa|TG], \\ & \text{graph}(T, Nds, R, TG). \end{array}$$

$$\begin{array}{ll} (B_{15}) \quad 1 : & \text{parents_of}([], _R, []) \\ (B_{16}) \quad \frac{R}{L} : & \text{parents_of}([PP|PPs], R, Pa) : - \\ & Pa = [PP|TPa], \\ & L, R : \text{parent_of}(PPs, R, TPa). \\ (B_{17}) \quad 1 - \frac{R}{L} : & \text{parents_of}([_PP|PPs], R, Pa) : - \\ & L, R : \text{parent_of}(PPs, R, Pa). \end{array}$$

Guard (G_2) sets up L to the number of possible parents under consideration (the number of all nodes minus 1) and R to the expected number of parents per family. Clause

(B_{15}) is the base case of the recursion while (B_{16}) adds a possible parent (PP) to the list of parents (Pa) and (B_{14}) eliminates the candidate. By setting the selection probability to R/L we expect R number of parents to be selected. The top part of Figure 1 shows the average number of family size from 1000 independent samples from this prior. There are two different values of n , the length of all nodes list, 5 and 10 for a single value of $r = 1$. The experimental results show that indeed the prior defines a normal distribution for the average family size with mean equal to 1. The bottom part of Figure 1 clearly shows that there is no bias in the selection of parents. The x-axis plots graph nodes and y-axis plots number of times the nodes were used as parents. Note that clauses B_{13} - B_{17} define a graph structure which might include cycles and as such not strictly a BN structure. We did so as to make the Dlp easier to follow and demonstrate the sampling distribution.

4.1 Likelihood based learning

Bayesian learning methods seek in the posterior distribution for either single models that maximise some measure or an approximation of the whole posterior. The posterior over models given some data $P(M|D)$ is proportional to the prior and a likelihood function, $P(M|D) \propto p(M)P(D|M)$. Since the space in all but trivial examples is too large to enumerate, various approximate methods have been introduced. Variational methods [10] approximate the inference on the evidence by considering a simpler inference task. Markov chain Monte Carlo algorithms sample from the posterior indirectly. So far we have concentrated on building priors that provide access to the choices via probabilistic paths. In this section we discuss one algorithm that can take advantage of the defined priors and its application to a real-world machine learning task.

4.2 Metropolis-Hastings

Metropolis-Hastings (MH) algorithms approximate the posterior by stochastic moves through the model space. A chain of visited model is constructed. At each iteration the last model added to the chain is used as a base from where a new model M' is proposed which is accepted or rejected stochastically. The distribution with which M' is reached from M is the proposal $q(M, M')$ and the acceptance probability is given by

$$\propto \frac{p(M'), P(M'|D), q(M, M')}{p(M), P(M|D), q(M', M)}$$

To our knowledge all MH algorithms in the literature, with the exception those based on SLPs, have distinct functions for computing the prior and the proposal. Standard MH requires two separate computations. The first is the prior over models: $p(M)$, and the second is a distribution for proposing a new model M' from current model M . The proposed model is accepted with probability that is proportional to the ratio given above which also includes the marginal likelihood of the model that measures the goodness of fit to the data $P(M|D)$. This often leads to restricting the choices of either the prior [8] or the proposal [4]. Furthermore, writing two programs that manipulate the same model space means that the algorithms are hard to extend to other spaces. The MH algorithm over Dlp requires the construction of a single program, that of the prior.

A generic MH algorithm for SLPs was suggested in [5] and further developed in [1]. The main idea is to use the choices in the probabilistic path as points from which alternative models can be sampled. Proposals are thus tightly coupled to the prior and take the form of a function f such that $\pi_j^M = f(\pi^M)$ where π^M is the path produced for deriving model M . π_j is the point from which M' will be sampled. As Dlp also provides a clear connection between computed instantiations and probabilistic choices the MH algorithm can be fitted over their priors. Open source software for MCMC simulations over the priors described here is web available ¹.

4.3 Chromatin interaction graph

We ran the MCMCMS system on the data from [16]. A simple prior was used that fits a gamma distribution over the parents with a mean value of 1.2 for the average family. The analysis presented in [16] was to build an 80% consensus graph based on repeated simulated annealing search using the Banjo software [15, 18]. The dataset consists of 4380 rows each representing a possible binding location for each of the 43 chromatin proteins (columns). We discretised the data as per the original paper by setting the strongest 5% of the bindings for each protein to 1 and the rest to 0.

We ran 10 chains of 5×10^5 iterations each. We then averaged the results in the form of a graph by including edges that appear more time than the threshold (80%). We will refer to the graph learnt by our method as M_{80} and to the graph in [16] as BN_{80} . In Fig.2 we show BN_{80} with edges that do not appear in M_{80} highlighted in blue, whereas Fig.3 shows M_{80} with edges that do not appear in BN_{80} highlighted in orange. Bidirectional edges depict averages that could only achieve the threshold for inclusion when considering both directions of the edge in the Markov chains. For ease of comparison both graphs are given with the same topology: that of [16]. Note, that this introduces artificial visual bias, making some new edges in M_{80} appear as long range effects. The directions of arrows in BN_{80} should be ignored as they are those of M_{80} except for those edges that are not present in M_{80} in which case the order is random.

There are 9 edges that were not in M_{80} , 6 edges that were not in BN_{80} and 40 common edges. M_{80} concurs with BN_{80} on all the major families. For instance the wiring of the classical heterochromatin family around HP1 is remarkably conserved. Graph M_{80} also contains all experimentally validated edges of BN_{80} . That is, the edges of HP1 with HP3 and HP6 and the edges of BRM with JRA, GAF and SU(VAR)3-7.

5 Conclusions

This paper discusses a general programming language for combining logical and probabilistic reasoning in logic programming specially for the purpose of defining prior Bayesian knowledge. The characterisation is of relevance not only to Dlp but also to other generative formalisms that combine logic and probability. We have argued that for certain classes of programs the kind of knowledge that can be represented in a convenient way is substantially improved. Furthermore, we illustrated via examples on how to write correct and efficient programs that capture knowledge from the Bayesian learning literature.

¹ <http://scibsf.bch.ed.ac.uk/nicos/sware/dlp/mcmcms/>

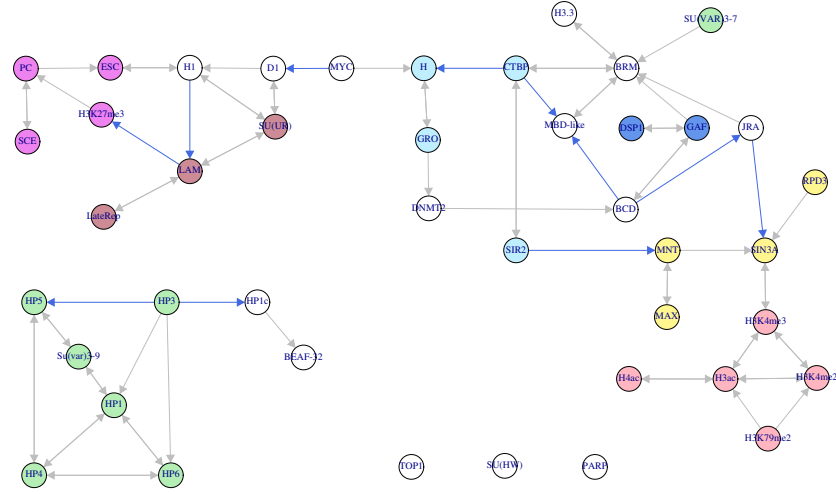


Fig. 2. Model BN_{80} , as presented in van Steensel et.al. (Ignoring edge directions.) Blue edges do not appear in M_{80} .

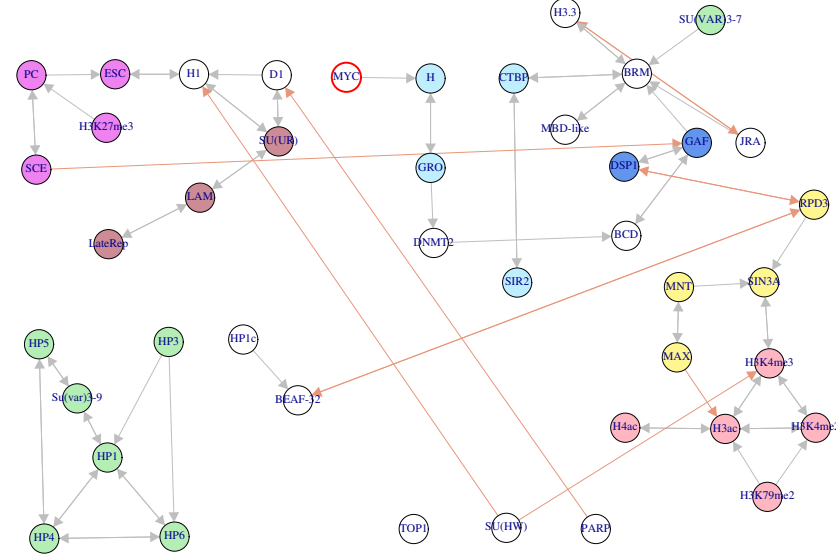


Fig. 3. Model M_{80} , learnt from 10 chains, each of 500,000 BNs. Shown interactions appear in 80% of the models. Single direction are for edges over the cut-off at one direction. Orange edges do not appear in BN_{80} .

We have used an MCMC schema over the probabilistic language to learn a graphical model from a recently published dataset. The consensus graph learnt by our method is in a

broad agreement ($\approx 80\%$) with a previously published graph. Furthermore the agreement coincides with the experimentally validated interactions.

References

1. Angelopoulos, N., Cussens, J.: MCMC using tree-based priors on model structure. In: UAI'01. pp. 16–23 (2001)
2. Angelopoulos, N., Gilbert, D.R.: A statistical view of probabilistic finite domains. In: Workshop on Quantitative Aspects of Programming Languages (2001)
3. Buntine, W.L.: Theory refinement on BNs. In: UAI'91. pp. 52–60 (1991)
4. Chipman H, George E, M.R.: Bayesian CART model search (with discussion). *Journal of the American Statistical Association* 93, 935–960 (1998)
5. Cussens, J.: Stochastic logic programs: Sampling, inference and applications. In: UAI'2000. pp. 115–122 (2000)
6. Cussens, J.: Parameter estimation in stochastic logic programs. *Machine Learning* 44(3), 245–271 (2001)
7. Di Pierro, A., Wiklicky, H.: An operational semantics for probabilistic concurrent constraint programming. In: *IEEE Comp.Soc.Conf. on Comp. Languages* (1998)
8. Friedman, N., Koller, D.: Being Bayesian about network structure. In: UAI-00. pp. 201–210 (2000)
9. Heckerman, D., Geiger, D., Chickering, D.: Learning BNs: The combination of knowledge and statistical data. *Machine Learning* 20(3), 197–243 (1995)
10. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. In: *Learning in Graphical Models*. MIT Press (1999)
11. Muggleton, S.: Stochastic logic programs. In: de Raedt, L. (ed.) *Advances in Inductive Logic Programming*, pp. 254–264. IOS Press (1996)
12. Ochs, M.F.: Knowledge-based data analysis comes of age. *Briefings in Bioinformatics* 11(1), 30–39 (2010), <http://bib.oxfordjournals.org/content/11/1/30.abstract>
13. Riezler, S.: Probabilistic Constraint Logic Programming. Ph.D. thesis, Neuphilologische Fakultät, Universität Tübingen, Tübingen, Germany (1998)
14. Sato, T., Kameya, Y.: Parameter learning of logic programs for symbolic-statistical modeling. *Journal of AI Research* 15, 391–454 (2001)
15. Smith, V., Jarvis, E., Hartemink, A.: evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* 18, S216S224 (2002), *intelligent Systems in Molecular Biology 2002 (ISMB02)*
16. van Steensel, B., Braunschweig, U., Filion, G.J., Chen, M., van Bemmelen, J.G., Ideker, T.: Bayesian network analysis of targeting interactions in chromatin. *Genome Research* 20(2), 190–200 (February 2010), <http://dx.doi.org/10.1101/gr.098822.109>
17. Werhli, A., Husmeier, D.: Reconstructing gene regulatory networks with bns by combining expression data with multiple sources of prior knowledge. *Stat. App. in Genetics and Molecular Biology* 6(1) (2007)
18. Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E.: Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 35943603 (December 2004)

Modeling Inter-practice Variation of Disease Interactions using Multilevel Bayesian Networks

Martijn Lappenschaar¹, Arjen Hommersom¹,
Stefan Visscher², and Peter J.F. Lucas¹

¹ Radboud University Nijmegen
Institute for Computing and Information Sciences
{mlappens,arjenh,peterl}@cs.ru.nl

² NIVEL (Netherlands Institute for Health Services Research)
S.Visscher@nivel.nl

Abstract. Multimorbidity is becoming a significant health-care problem for western societies, especially within the elderly. Since medical knowledge is mostly organized around single diseases, it is unlikely that the elderly patient with multiple diseases receives appropriate treatment. To get a grip on complex interactions, we aim to model domains using hierarchies, for example, patient characteristics, pathophysiology, symptomatology and treatment. For this we introduce *Multilevel Bayesian networks*, which we have applied to clinical data from family practices in the Netherlands on heart failure and diabetes mellitus. We compare the outcomes to conventional methods, which reveals a better insight of interactions between multiple diseases.

1 Introduction

Recent epidemiological research in the Netherlands indicates that more than two third of all patients older than 65 years have two or more chronic diseases at the same time; this problem, one of the most challenging of modern medicine, is referred to as the problem of comorbidity or multimorbidity. Where *comorbidity* is defined in relation to a specific index condition, the term *multimorbidity* has been introduced in chronic disease epidemiology to refer to any co-occurrence of two, but often more than two, multiple chronic or acute diseases within a person. The introduction of this term indicated a shift of interest from a given index disease (i.e. the primary disease) to individuals having multiple diseases.

There is no guarantee that, in case of a patient with multiple diseases, treating each disease individually is optimal. The need of an integrated optimal treatment for a patient with multiple diseases also implies the need for an integrated research methodology of multiple diseases. However, medical researchers often focus on an index disease rather than looking at multimorbidity in total.

Typically, regression methods are used to analyze the variance in disease variables, where researchers focus on the power of specific variables for predicting the presence or absence of specific diseases [22]. Where linear regression is used

for continuous outcome variables, logistic regression is mostly used for dichotomous outcome variables. In case patients can be divided into groups, *multilevel* regression can be used to analyze the group dependent variance by adding extra variance components [7].

In contrast to using regression of fixed functional form, the patient data can also be modeled using probabilistic graphical models, such as Bayesian networks [11]. The edges of the graphical model then represent relationships between patient characteristics, pathophysiology and diagnostic tests for the disease of interest, which naturally generalizes to multiple diseases. However, multilevel modeling has not been studied in this context.

In this paper we introduce a new representation of multilevel disease models using Bayesian networks – which we call *multilevel Bayesian networks* – of which the multilevel regression model is a special case. This gives us the advantage that multiple models, e.g. of diseases, can be merged into one model, which allows examination of the interactions between them. Moreover, we apply this framework to patient data from family practices in the Netherlands. Its effectiveness is shown by comparing the model to the traditional methods based on regression analysis.

2 Multimorbidity: Context and Related Research

An Abstract Disease Model The context of multimorbidity is illustrated by Fig. 1 which provides an abstract view on the problem. The left-hand side shows the typical relationships between variables when considering a single disease. They form a hierarchical topology: genetics and environment, patient characteristics, disease, pathophysiology, and measurable variables, i.e. specific signs, symptoms and laboratory results.

If we represent multiple diseases in the same model, all kinds of interaction between variables within this model can be identified, as is illustrated at the right-hand side of Fig. 1. Mutual dependences between the two diseases may

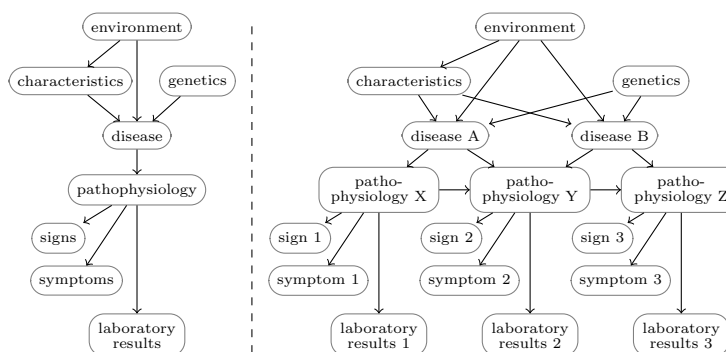


Fig. 1. Abstract model of a single disease (left) and multiple diseases (right).

concern their pathophysiology, symptoms, signs, and lab results. By modeling these interactions explicitly, better decisions can be made for patients having multiple diseases. Moreover, when considered separately, single disease models often contain a lot of overlap with each other, which may be avoided by integrating different disease models into a single model.

Normally, in scientific research, one would investigate diseases separately, resulting in different predictive values of variables shared by both diseases. Recently, multilevel regression analysis was used to investigate the influence of particular family practice variables on hypertension and diabetes mellitus, revealing an inter-practice variance in predictability [10]. However, since interactions could have an additive effect on prevalence, this yields no insight into the predictive value in case both diseases are present. Actually, we need an extra regression on the combined diagnosis to be able to conclude on such information.

In regression methods the variance of the observations is minimized with respect to the dependency between variables. Multilevel analysis also tries to explain the variance caused by grouping variables that intermediate on the lower level variables, i.e. it allows the intercept and slope, that determines the linear dependency between two variables, to alter for different groups. To analyze complex multimorbidity models one might have to deal with large datasets in which many variance is introduced. This can be due to the fact that data is collected from different kind of sources (e.g. family practices) or the data represents patients of all kind of populations (social, economic, and demographic differences). If we would ignore this, identifying interactions between disease variables such as pathophysiology and laboratory results could be difficult or erroneous.

Ultimately, we need one model that explains, both the variance, introduced in the observations, and the interactions in case of multiple diseases. If we can translate the multilevel regression models, which can deal with the variance explained by hierarchical structures themselves, to a graphical representation in such a way that we are able to connect multiple models of different diseases together, we also make them dependent on the interactions between diseases.

Related Research Much of the medical research relies on regression models which are applied to a single disease, and, thus, ignore the complexity of multimorbidity. Prevalence of multimorbidity are studied in family practices [1], sometimes with clustering of specific diseases [8]. These results illustrate the impact and complexity, but give little insight into interactions between diseases.

More advanced methods to analyze multimorbidity in particular were not available until recently. A network analysis of pairwise comorbidity correlations for 10.000 diseases from 30 million medical records illustrated the complexity of many physiological processes in a context of patient characteristics such as ethnicity and genetic predisposition [6]. Markov blanket models and Bayesian model averaging were used in algorithms for learning patient-specific models from clinical data, to predict the outcome of sepsis or death in case of cardiovascular diseases [21]. To deal with polypharmacy, there is recent work of a Bayesian network meta-analysis to compare antihypertensive treatments in randomized controlled trials [16]. The method allows a comparison of multiple treatments

where only a subset of treatments were compared in each trial. This mixed treatment comparison was facilitated with a framework of Markov models to be able to monitor disease progression [13].

Bayesian graphical modeling [18] is presented as a framework for generalized linear models, including multilevel and hierarchical models, with the aim to represent the conditional independence assumptions for parameters and observables and to make them the basis for a local computational strategy, generally based on Markov Chain Monte Carlo (MCMC) methods. It addresses solutions to deal with overdispersion, hierarchical modeling, dependency between coefficients, model criticism, making predictions, covariates and missing populations.

Although not specially designed for multimorbidity, similarity networks and Bayesian multinets [2] may offer a suitable method to represent uncertain knowledge in case of multiple diseases. An advantage of these methods is the possibility to represent asymmetric independence assertions, meaning that dependency between variables may only occur for certain values of these variables.

In the next section, basic techniques used in this paper are briefly reviewed.

3 Preliminaries

In this section we provide the basic concepts that we will use when modeling multimorbidity. Before moving on to the regression methods and Bayesian networks we first summarize basic elements of probability theory putting emphasis on multivariate probability distributions. Further on, we will discuss the issues that need to be dealt with when modeling multiple diseases.

3.1 Probability Theory

The patients's characteristics, pathophysiology, investigations, etc., can be seen as random variables each with its own distribution. Formally, random variables are denoted with uppercases, and observations with lowercases. We assume there is some joint, or multivariate, probability distribution over the set of random variables X , denoted by $P(X)$. The probability of a conjunction of two sets of variables, $X \wedge Y$, is denoted as $P(X \wedge Y)$ and also as $P(X, Y)$. The marginal distribution of $Y \subseteq X$ is then given by summing (or integrating) over all the remaining variables, i.e., $P(Y) = \sum_{Z=X \setminus Y} P(Y, Z)$. A conditional probability distribution $P(X | Y)$ is defined as $P(X, Y)/P(Y)$. Two variables X and Y are said to be conditionally independent given a third variable, Z , if $P(X | Y, Z) = P(X | Z)$.

In case a variable X is discrete, the variable is bounded by a finite set of possible values x , a probability is then denoted by $P(X = x)$. In case the outcome space of a variable X is the set of real numbers \mathbb{R} or a subset thereof, one uses the probability $P(X \leq x)$.

3.2 Linear Regression

In general, in medical research, we have a dataset of observations of a number of patients, and we could see them as possible outcomes of the random variables. The variables can be split up in several domains. We can distinguish outcome variables, denoted as O_i , and explanatory variables, denoted as E_i . Some explanatory variables act on a group level, i.e. they have the same value for each individual within a certain group, which are denoted as L_i .

Linear regression tries to fit the observations of a random continuous variable (assuming it is normally distributed) into a linear model. This is done using an algorithm, e.g. a least square method, that minimizes the defiance of the observations with respect to the model parameters (the variance). Typically, we want to explain an observation o with respect to explanations e_i assuming that the observations o are possible outcomes of a random variable O . If the vectors $(1, e_1, \dots, e_i, \dots, e_n)^T$ are explanations, linear regression often estimates the parameters $\beta = (\beta_0, \beta_1, \dots, \beta_i, \dots, \beta_n)^T$, such that

$$P(O \mid e) \sim \mathcal{N}(\mu, \Sigma), \text{ with } \mu = \beta^T e \quad (1)$$

for every explanation e . Linear regression only makes sense in case of continuous variables. In case of disease variables this mostly only accounts for physical measurements. For example, a linear relation between two different kind of blood measurements BM_1 and BM_2 , e.g. the low density lipoprotein (LDL) and high density lipoprotein (HDL) blood values, could be modeled as:

$$P(BM_1 \mid BM_2 = bm_2) \sim \mathcal{N}(\beta_0 + \beta_1 bm_2, \Sigma)$$

For more details about linear regression and other regression methods, especially in the medical area, one is referred to [22].

3.3 Multilevel Regression

In multilevel regression, part of the variance is explained due to group effects, i.e. the intercept and slope of the linear dependencies is allowed to alter amongst different groups. Now suppose we have a set of observations l_j , with $1 \leq j \leq m$, that have the same value within a certain group of patients, and based on that we can divide the patients into k groups. We could simply add these variables to the regression model as extra predictors. If we have $e = (1, e_1, \dots, e_n, l_1, \dots, l_m)^T$ as possible multivariate outcome, and β as $(\beta_0, \beta_1, \dots, \beta_n, \beta_{n+1}, \dots, \beta_{n+m})^T$ we keep a model as defined in Equation (1), having $n + m + 1$ degrees of freedom.

Multilevel regression, however, offers a different approach. For each k^{th} group we define a linear regression model, with O_k as random outcome variable, and allow dependency of the regression coefficients on the variables l_j and certain deviation from the overall mean. With $e = (1, e_1, \dots, e_n)^T$, $l = (1, l_1, \dots, l_m)^T$, $\beta_k = (\beta_{k0}, \dots, \beta_{kn})^T$, $\delta_k = (\delta_{k0}, \dots, \delta_{kn})^T$, Γ_k a matrix consisting of γ_{ij}^k , and $\delta_{ki} \sim \mathcal{N}(0, \Sigma_\delta)$, the model becomes:

$$P(O_k \mid e, g) \sim \mathcal{N}(\mu, \Sigma), \text{ with } \mu = (\delta_k + \Gamma_k g)^T e \quad (2)$$

The model is now more complex and the number of degrees of freedom is $k(n+1)(m+2)$. For example, if we extend our previous example by grouping on gender represented by a variable *gen*, and allow an influence of gender on the relation between the two blood measurements, the model then becomes:

$$\begin{aligned} P(BM_1 \mid \text{male}, bm_2) &= \mathcal{N}(\delta_0^m + \gamma_0^m + (\delta_1^m + \gamma_1^m)bm_2, \Sigma) \\ P(BM_1 \mid \text{female}, bm_2) &= \mathcal{N}(\delta_0^f + \gamma_0^f + (\delta_1^f + \gamma_1^f)bm_2, \Sigma) \end{aligned}$$

The parameters of multilevel regression models can be estimated using an restricted iterative generalized least square (RIGLS) method, which coincides with restricted maximum likelihood (REML) in Gaussian models [3]. It estimates the parameters by alternating the optimizing process between the fixed parameters (γ_{kij}) and the stochastic parameters (δ_{ki}) until convergence is reached, and is equivalent to the maximum likelihood estimation in standard regression.

3.4 Generalized Regression Models

The former model assumes that the random outcome variable O is normally distributed. But suppose we want to consider a dichotomous outcome variable with only the possible values 'yes' and 'no'. An approach to deal with non-normally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution explicitly into the model. This class of statistical models are called generalized linear models. They are defined by three components: an outcome variable O that has an expected value $E[O|e]$, a linear additive regression equation that produces an unobserved (latent) predictor η of the outcome variable O , and a link function that links the expected values of the outcome variable O to the predicted values for η . In logistic regression the link function is given by $\eta = \text{logit}(E[O|e]) = \log \frac{E[O|e]}{1-E[O|e]}$. The logistic multilevel model then becomes:

$$\text{logit}(E[O_k] \mid e, l) = (\delta_k + \Gamma_k l)^T e$$

The conditional probability in case of logistic regression is then defined as:

$$P(O_k \mid e, l) = \frac{1}{1 + e^{-x}}, \text{ with } x = (\delta_k + \Gamma_k l)^T e \quad (3)$$

For example, we are interested in the predictive value of blood measurement BM_1 and BM_2 to an dichotomous outcome variable such as disease $D1$ with possible values 'yes' or 'no'. The multilevel logistic regression model with respect to gender then becomes:

$$\begin{aligned} \text{logit}(E[D_1] \mid \text{male}, bm_1, bm_2) &= \delta_0^m + \gamma_0^m + (\delta_1^m + \gamma_1^m)bm_1 + (\delta_2^m + \gamma_2^m)bm_2 \\ \text{logit}(E[D_1] \mid \text{female}, bm_1, bm_2) &= \delta_0^f + \gamma_0^f + (\delta_1^f + \gamma_1^f)bm_1 + (\delta_2^f + \gamma_2^f)bm_2 \end{aligned}$$

Parameters for dichotomous outcomes are estimated with marginal and penalized quasi-likelihood (MQL/PQL) algorithms [4]. Alternatively MCMC methods such as Gibbs Sampling can be used [15].

3.5 Bayesian Networks

Bayesian networks offer an effective framework for knowledge representation and reasoning under uncertainty [11]. Formally, a *Bayesian network*, or BN, is a tuple $\mathcal{B} = (G, X, P)$, with $G = (V, E)$ a directed acyclic graph (DAG), $X = \{X_v \mid v \in V\}$ a set of random variables indexed by V , and P a joint probability distribution. X is a Bayesian network with respect to the graph G if P can be written as a product of the probability of each random variable, conditional on their parent variables:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{v \in V} P(X_v = x_v \mid X_j = x_j \text{ for all } j \in \pi(v)) \quad (4)$$

where $\pi(v)$ is the set of parents of v (i.e. those vertices pointing directly to v via a single arc). If there are continuous variables, the definition is similar, and can be defined by using the probability density function. While the conditional probabilities could be estimated using regression methods [14], parameter and structure learning methods for Bayesian networks are readily available [9].

For example, suppose we have two binary variables, D for disease present yes/no and G for gender, both having an direct effect on the blood measurements BM_1 and BM_2 . Besides that BM_1 affects BM_2 also directly. Using marginalization, we obtain the probability for BM_2 by:

$$P(BM_2) = \sum_D \sum_G \sum_{BM_1} P(BM_2 \mid BM_1, D, G) P(BM_1 \mid D, G) P(D) P(G)$$

Conditional independence in Bayesian networks is an important concept when modeling multimorbidity. When considering three vertices u , v and w we can distinguish certain types of dependencies:

- v is a tail-tail vertex ($u \leftarrow v \rightarrow w$)
- v is a head-tail vertex ($u \rightarrow v \rightarrow w$)
- v is a head-head vertex ($u \rightarrow v \leftarrow w$)

For the first two situations we obtain independence between X_u and X_w if we condition on X_v , i.e. $P(X_u \mid X_w, X_v) = P(X_u \mid X_v)$, also denoted as $X_u \perp\!\!\!\perp X_w \mid X_v$, whereas $X_u \not\perp\!\!\!\perp X_w \mid \emptyset$. In the third situation, the situation is reversed, as X_u and X_w are unconditionally independent, whereas they become dependent when conditioning on X_v , i.e., $X_u \perp\!\!\!\perp X_w \mid \emptyset$ and $X_u \not\perp\!\!\!\perp X_w \mid X_v$. The Markov blanket (MB) of a vertex contains all the variables that shield the vertex from the rest of the network, meaning that if all variables within the MB can be observed, this is the only knowledge needed to predict the behavior of that vertex [11].

It is appealing to define disease variables as binary variables, i.e. the disease of interest is present yes or no. Socio-economical and demographic variables are often categorical (sometimes ordered) or numerical (e.g. age). Laboratory investigations are often continuous (especially blood measurements), but can be discretized, e.g. blood glucose levels could be defined as normal, subclinical, and clinical. Variables can be dependent on each other, or independent, which

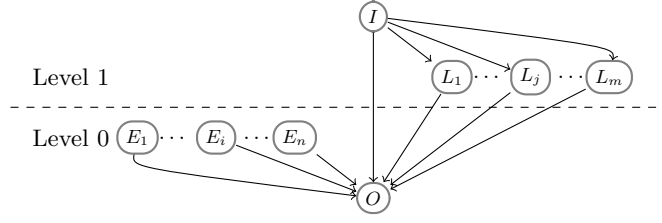


Fig. 2. Bayesian network of multilevel regression.

can be represented in a Bayesian network, defining variables as vertices and dependencies as edges. If the structure is unknown it can be learned.

Ideally, we expect to obtain some kind of hierarchical topology in the learned structure, just as described in Fig 1. In fact, we can put restrictions into the learning algorithm to force such a topology. If we consider the disease variables, an association might be present between them, but there's a chance we could make them conditional independent if we observe the environmental and patient's characteristics variables, i.e. they serve as tail-tail variables with respect to disease variables. At the other hand looking at the laboratory results, those might act as head-head variables with respect to diseases, therefore we cannot make the diseases conditional independent looking solely to laboratory results.

4 Multilevel Bayesian Networks

In this section, we introduce the multilevel Bayesian network (MBN) formalism as interpretation of multilevel regression. First, we briefly explore the relation between multilevel regression models and Bayesian networks. Then, we generalize this by allowing more structure within the model. We discuss the building and learning of such models and compare this to the regression approach.

4.1 Multilevel Regression Analysis as a Bayesian Network

In multilevel regression, the outcome variable O depends on the explanations $e = e_1, \dots, e_n$ and $l = l_1, \dots, l_m$. In a Bayesian network approach, we model O as a conditional probability distribution given the set of parents E_1, \dots, E_n , L_1, \dots, L_m , and I_1, \dots, I_k , i.e., we now interpret the explanations and group explanations as instantiations of random variables. The variables I_j , with $1 \leq j \leq k$ is an indicator variable for grouping of objects at a certain level j . Fig. 2 then shows the corresponding Bayesian network, assuming independent predictors.

Clearly, this model is not realistic in case of multimorbidity domains. There is no structure present between predictors and we have only one outcome variable of interest. In general, the opposite is more likely to be true, i.e. multiple outcome variables, multiple dependencies between predictors and variables that are both predictor and outcome variable. While discriminative learning algorithms, such as regression, are good for prediction, they do not provide insight

into the domain, nor can they be used to model interactions necessary in the case of multimorbidity. Bayesian networks have the ability to give such insight, by allowing dependencies between variables.

4.2 Multilevel Bayesian Networks in General

The idea of a multilevel Bayesian network is that I variables split the domain into different categories with a deterministic effect on variables that are constant within a category (L). Some variables (E) are group-independent, though structure may exist between these variables. Other variables (O) depend both on grouping and other variables in the same or higher levels. The Bayesian network is constrained in the sense that no edges exist from a lower-level variable to a higher-level variable. This ensures that we keep the hierarchical structure obtained with multilevel regression methods (see Fig. 3). A *multilevel Bayesian network* is defined as a tuple (G, N, I, E, O, L, P) such that:

- (G, V, P) , where $V = I \cup E \cup O \cup L$, forms a Bayesian network;
- I , E , O , and L are pair-wise disjoint;
- $N \in \mathbb{N}$ denotes the number of levels on top of the base level 0;
- $I = \{I_1, \dots, I_N\}$ are variables such that each value of such a variable contains a group. It holds that I_j is the only parent of I_{j-1} in G for all $1 \leq j \leq N$;
- $E = \{E^0, \dots, E^N\}$ where each E^j is a set of variables corresponding to level j , such that if $(V \rightarrow E_i^j) \in G$, then $V \in (E \cup O)^{j+k}$, with $k \geq 0$;
- $O = \{O^0, \dots, O^N\}$ where each O^j is a set of variables corresponding to level j , such that if $(V \rightarrow O_i^j) \in G$, then $V \in (E \cup O)^{j+k}$ or $V \in L^{j+k+1}$ $k \geq 0$;
- $L = \{L^1, \dots, L^N\}$ where each L^j is a set of group variables corresponding to a level j . If $L_i^j \in L^j$ is a group variable, then it holds that (i) $(I_j \rightarrow L_i^j) \in G$, (ii) there are no other variables V such that $V \rightarrow L_i^j$, and (iii) $P(L_i^j | I_j)$ is deterministic.

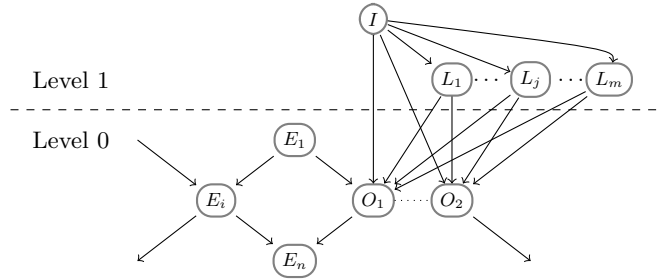


Fig. 3. Bayesian network of multilevel regression with an improved structure between predictors (E_i) and outcome variables (O_i).

4.3 Building Multilevel Bayesian Networks

In order to build the structure between intra-level variables, we can make use of two approaches. We can either model the structure manually based on existing medical knowledge or learn the structure from data. Structure learning of Bayesian networks offers a suitable method to learn these dependencies. The constraints imposed by the multilevel Bayesian network can be captured by blacklisting and whitelisting edges, which can be incorporated into a wide range of structure learning algorithms (see e.g., [17]). For example, the necessary edges between I and other variables are whitelisted, whereas edges from a lower level to a higher level are all blacklisted. The parameters can be learned using standard Bayesian network techniques. Compared to multilevel regression models, it is also possible to use a Bayesian approach for learning the parameters [19] and therefore include even more domain knowledge to the model.

Model validation can be achieved by using information criteria such as the Deviance Information Criteria and the Bayesian Information Criteria [20]. Alternatively, standard cross validation (e.g. k-fold cross validation) is a robust method to validate regression and Bayesian models [12].

5 Modeling Inter-practice Variation

5.1 Problem domain and data

In this paper, we apply the MBN approach for modeling inter-practice variations for predicting heart failure and diabetes mellitus. Data was collected by the Netherlands Information network of General Practice (LINH). In 1996, they started as a register of referrals of general practitioners to medical specialists. Information about contacts and diagnoses, prescriptions, referrals and laboratory and physiological measurements are extracted from the information systems. The LINH database contains information of routinely collected data from approximately 90 general practices. Unless patients moved from practices, and practices opted out, longitudinal data of approximately 300.000 distinct patients are stored. Patients under 25 were excluded, because of their low probability on multimorbidity. Practices who recorded during less than six month were also excluded from statistical analysis. Eventually, we used data of 218333 patients. Morbidity data were derived from diagnoses, using the international classification of primary care (ICPC) and anatomical therapeutic chemical (ATC) codes.

5.2 Results

Our main variables of interest are heart failure and diabetes mellitus. The predictors are shown in Table 1, with the urbanity of the practice's area as higher level variable. Multilevel logistic regression leaves us then with five separate models, for which the parameters are estimated using MLWin [5]. To obtain the parameters of the MBN interpretation we ran a MCMC method which is available in the WinBUGS software [19]. Dichotomous variables are modeled using

| | RIGLS | | MCMC-fixed | | MCMC-learn | |
|--|---------------|-------|---------------|-------|---------------|-------|
| Diabetes Mellitus | β | Odds | β | Odds | β | Odds |
| Intercept | -5.800 (0.3%) | | -5.678 (0.3%) | | -5.866 (0.3%) | |
| Age | 0.029 | 1.029 | 0.028 | 1.028 | 0.063 | 1.065 |
| Gender (ref = male) | -0.090 | 0.914 | -0.089 | 0.915 | -0.222 | 0.801 |
| Overweight/obesity | 0.545 | 1.725 | 0.513 | 1.671 | 1.189 | 3.282 |
| Lipid disorder | 1.862 | 6.437 | 1.855 | 6.392 | | |
| Hypertension | 1.736 | 5.675 | 1.758 | 5.800 | | |
| Atherosclerosis | -0.047 | 0.954 | -0.052 | 0.949 | | |
| Heart failure | 0.124 | 1.132 | 0.178 | 1.194 | | |
| Retinopathy | 2.225 | 9.253 | 2.269 | 9.669 | | |
| Angina pectoris | -0.387 | 0.679 | -0.409 | 0.665 | | |
| Stroke / CVA | -0.262 | 0.770 | -0.269 | 0.766 | | |
| Renal disease | 0.162 | 1.176 | 0.183 | 1.200 | | |
| Cardiovasc. symptoms | -0.165 | 0.848 | -0.163 | 0.850 | | |
| Urbanity (ref=urban) | | | | | | |
| strongly urban | 0.232 | 1.261 | 0.243 | 1.275 | -0.326 | 0.722 |
| modestly urban | 0.390 | 1.477 | 0.399 | 1.490 | 0.389 | 1.476 |
| little urban | 0.362 | 1.436 | 0.342 | 1.408 | -0.934 | 0.393 |
| not urban | 0.388 | 1.474 | 0.230 | 1.259 | -0.313 | 0.731 |
| Model validation average accuracy (cross validation) | | 89% | | 89% | | 88% |

| Heart Failure | β | Odds | β | Odds | β | Odds |
|--|----------------|-------|---------------|-------|---------------|-------|
| Intercept | -11.373 (0.0%) | | -11.20 (0.0%) | | -11.24 (0.0%) | |
| Age | 0.101 | 1.106 | 0.101 | 1.106 | 0.105 | 1.111 |
| Gender (ref=male) | -0.195 | 0.823 | -0.204 | 0.815 | -0.160 | 0.852 |
| Overweight/obesity | 0.524 | 1.689 | 0.470 | 1.600 | | |
| Diabetes mellitus | 0.228 | 1.256 | 0.726 | 2.067 | 0.330 | 1.391 |
| Lipid disorder | 0.159 | 1.172 | -0.832 | 0.435 | | |
| Hypertension | 0.728 | 2.071 | 0.425 | 1.530 | 0.963 | 2.618 |
| Atherosclerosis | 0.482 | 1.619 | 0.231 | 1.260 | 0.655 | 1.925 |
| Retinopathy | 0.270 | 1.310 | 0.099 | 1.104 | | |
| Angina pectoris | 0.795 | 2.214 | 0.781 | 2.184 | | |
| Stroke / CVA | 0.328 | 1.388 | 0.334 | 1.397 | | |
| Renal disease | 0.630 | 1.878 | 0.632 | 1.881 | 0.720 | 2.054 |
| Cardiovasc. symptoms | 0.954 | 2.596 | 0.969 | 2.636 | | |
| Urbanity (ref=urban) | | | | | | |
| strongly urban | 0.135 | 1.145 | 0.147 | 1.158 | | |
| modestly urban | 0.166 | 1.181 | 0.176 | 1.192 | | |
| little urban | 0.352 | 1.422 | 0.375 | 1.456 | | |
| not urban | 0.289 | 1.335 | 0.276 | 1.318 | | |
| Model validation average accuracy (cross validation) | | 89% | | 89% | | 95% |

Table 1. Parameter estimations and cross validation of multilevel analysis for Heart Failure and Diabetes Mellitus. RIGLS = restrictive iterative general least square method for the multilevel logistic regression model, MCMC-fixed = Monte Carlo Markov Chain method for the multilevel Bayesian network without structure learning, MCMC-learn = same as MCMC-fixed but with structure learning.

a Bernoulli distribution. Parameter estimates and the average accuracy of predicting heart failure and diabetes mellitus are presented in in Table 1. Validated using a 10-fold validation, the table shows that the MBN model is in line with results obtained by multilevel regression.

The next step is structure learning of predictors and outcome variables while maintaining the multilevel structure as mentioned in Section 4. Diseases are obviously not a cause of practice and patient characteristics such as age and gender. The *bnlearn* package [17] in the statistical software R provides both constraint based as scoring algorithms to learn the structure of a Bayesian network. The score based methods reveal the most appealing structure for the available data. See Fig. 4a for the resulting Bayesian network structure.

Some of the directions of certain edges is opposite to what the domain experts would expect, e.g. angina pectoris and heart failure is pointing towards atherosclerosis, but in reality the latter is seen as a cause of the first and not a

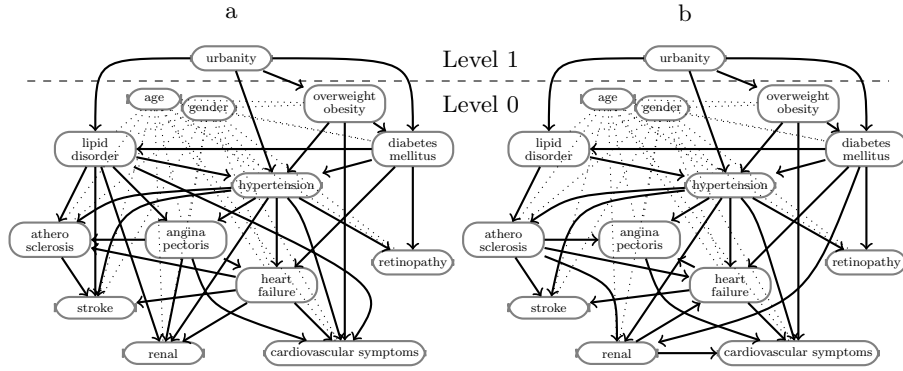


Fig. 4. Structure learning of Heart Failure and Diabetes in Family Practices, (a) with inter-level restrictions only, and (b) with intra-level restrictions (expert opinions / evidence from other research) as well.

| | obesity /overw. | diabetes mellitus | lipid disorder | hyper- tension | athero- sclerosis | renal disease | heart failure | angina pectoris | stroke /CVA | retino- pathy |
|------------|--------------------|----------------------|-------------------|-------------------|----------------------|------------------|------------------|--------------------|----------------|------------------|
| Learned | 89% | 88% | 84% | 86% | 96% | 95% | 95% | 94% | 96% | 97% |
| Restricted | 89% | 88% | 85% | 87% | 96% | 95% | 95% | 95% | 96% | 97% |

Table 2. Accuracy for predicting diseases in a multilevel Bayesian network for the model containing both heart failure and diabetes mellitus. The 'Learned' model corresponds with Fig 4a, and the 'Restricted' model corresponds with Fig 4b.

consequence. Probably, this opposite direction is due to the fact that atherosclerosis is mostly diagnosed clinically by interpreting symptoms and signs of the disease. By incorporating domain knowledge into the model and re-running the structure learning algorithm, we obtain the model as shown in Fig. 4b.

We have learned the probability distributions of both the learned and restricted model. Using a 10-fold cross validation we calculated the accuracy of predicting not only diabetes mellitus and heart failure, but also for the other diseases present in the MBN. See Table 1 and 2 for an overview of the results. The accuracy of predicting diabetes mellitus is similar to the previous models, whereas the accuracy of predicting heart failure is 6% better. The accuracies for the other disease variables, ranging between 84% and 97%, are slightly better in the restricted model. Looking into interactions between predictors, the structured model is more accurate when looking to heart failure. Table 3 shows the estimated and true prevalences of heart failure in the presence of multiple comorbidities. The estimations of the structured model are closer to the actual data. Clearly, the problem with the regression model is that it does not recognize the fact that the prevalence of heart failure is independent of obesity when conditioned on hypertension and diabetes mellitus.

| | obesity | | | | no obesity | | | |
|-----------------------|---------------|------------------|---------------|------------------|---------------|------------------|---------------|------------------|
| | diabetes | | no diabetes | | diabetes | | no diabetes | |
| | hyper-tension | no hyper-tension | hyper-tension | no hyper-tension | hyper-tension | no hyper-tension | hyper-tension | no hyper-tension |
| Heart Failure | | | | | | | | |
| Multilevel Regression | 10 | 5.1 | 8.1 | 4.1 | 6.5 | 3.2 | 5.2 | 2.6 |
| Multilevel Network | 9.1 | 0.0 | 4.1 | 0.7 | 9.3 | 0.3 | 5.2 | 0.6 |
| Calculated from data | 10 | 0.0 | 4.3 | 0.7 | 10.3 | 0.3 | 5.5 | 0.6 |

Table 3. Prevalences (in percentages) of heart failure in the presence of obesity, diabetes and hypertension, based on model parameters compared to actual values.

6 Discussion

In this paper we introduced Bayesian networks as an interpretation of multilevel analysis. Using patient data from family practices, its predictive value for heart failure and diabetes mellitus is just as good compared to traditional methods such as multilevel regression analysis, despite a reduced number of predictors.

The advantage of multilevel Bayesian networks is that it allows multiple outcome variables within one model, reducing redundancy of multiple multilevel regression models. Furthermore, we can add intra-level structures between variables giving extra insight into dependencies. Bayesian networks can be used to model conditional independence between variables, as we have seen with heart failure. We could perform a complete structure learning of the data, ignoring the hierarchy of variables.

But in practice the model is then very prone to assign causality from lower level variables to higher level variables, which in fact is not possible if we define the hierarchy properly. In case of patient data within a multimorbidity setting it is appealing to use the hierarchical topology: genes and environment (e.g. urbanity) – patient characteristics (e.g. age, gender, habits) – pathophysiology (diseases, syndromes) – and symptomatology (e.g. symptoms, signs, laboratory results), which can be modeled well using multilevel Bayesian networks.

The disadvantage of using patient data from family practices, is that it is driven by the actions of the physician. Since most diabetic patients are sent to an eye doctor, obviously we will find a overestimated relation between diabetes and retinopathy. Data of specific pathophysiologic tests are not available (yet), so diagnoses are not strict but act with a certain probability depending on the specificity of the tools used in family practice. Future work will also focus on patient data retrieved from randomized controlled trials, with the additional difficulty to learn multiple parameters using low patient counts.

Furthermore, since the data available will never provide a full causal model, it is important to make use of expert input. Besides putting restrictions on existing variables, one might also introduce variables that are missing from the data, but which may add crucial explanatory power. This is possible in BNs, and thus MBNs can also use the same expertise to quantify the probabilistic relationships involving these missing variables even though no data exists for them. So, multilevel Bayesian networks enforces a kind of supervised structural learning with respect to variance explained by higher level variables.

References

1. M Fortin, et al. Prevalence estimates of multimorbidity: a comparative study of two sources. *BMC Health Services Research* 10:111, 2010.
2. D Geiger, D Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence* 82:45-76, 1996.
3. H Goldstein. Restricted unbiased iterative generalised least squares estimation. *Biometrika* 76:622-623, 1989.
4. H Goldstein, J Rabash. Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society (Series A)* 159:505-512, 1996.
5. H Goldstein, W Browne, J Rasbash. Multilevel modeling of medical data. *Statistics in Medicine* 21(21):3291-3315, 2002.
6. CA Hidalgo, N Blumm, AL Barabasi, NA Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol.* 2009;5:e1000353.
7. JJ Hox. *Multilevel Analysis: techniques and applications*, 2nd ed. Routledge 2010.
8. A Marengoni, D Rizzuto, HX Wang, B Winblad, and L Fratiglioni. Patterns of chronic multimorbidity in the elderly population. *J Am Geriatr Soc* 57:225-230, 2009.
9. RE Neapolitan. *Learning Bayesian networks*. Prentice Hall, 2004.
10. MMJ Nielen, et al. Inter-practice variation in diagnosing hypertension and diabetes mellitus: a cross-sectional study in general practice. *BMC Fam Pract* 10:1-6, 2009.
11. J Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
12. R Picard, D Cook. Cross-validation of Regression Models. *Journal of the American Statistical Association* 79(387):575-583, 1984.
13. MJ Price, NJ Welton, AE Ades. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Statistics in Medicine* 30:140-151, 2011.
14. F Rijmen. Bayesian networks with a logistic regression model for the conditional probabilities. *Int. J. Approx. Reasoning* 48(2):659-666, 2008.
15. MH Seltzer, et al. Bayesian analysis in applications of hierarchical models: issues and methods. *Journal of Educational and Behavioral Statistics* 21:131-167, 1996.
16. S Sciarretta, F Palano, G Tocci, R Baldini, M Volpe. Antihypertensive treatment and development of heart failure in hypertension. *Arch Intern Med.* 171:384-394, 2011.
17. M Scutari. *Learning Bayesian Networks with the bnlearn R Package*. *Journal of Statistical Software* 35(3):122, 2010.
18. DJ Spiegelhalter. Bayesian Graphical Modelling: A Case-Study in Monitoring Health Outcomes. *Applied Statistics* 47-1:115-133, 1998.
19. DJ Spiegelhalter, A Thomas, N Best, D Lunn. *WinBUGS User Manual; Version 1.4*. MRC Biostatistics Unit, Cambridge UK, 2001.
20. DJ Spiegelhalter, NG Best, BP Carlin, A van der Linde. Bayesian measures of model complexity and fit. *J.R. Statist. Soc. B* 64-4:583-639, 2002.
21. S Visweswaran, DC Angus, M Hsieh, L Weissfeld, D Yealy, GF Cooper. Learning patient-specific predictive models from clinical data. *J Biom Inform.* 43:669-85, 2010.
22. E Vittinghoff, DV Glidden, SC Shiboski, CE McCulloch. *Regression Methods in Biostatistics: linear, logistic, survival and repeated measures models*. Springer 2005.

Towards a Method of Building Causal Bayesian Networks for Prognostic Decision Support

Barbaros Yet¹, Zane Perkins², William Marsh¹, Norman Fenton¹

¹School of Electronic Engineering and Computer Science, Queen Mary, University of London,
Mile End Road, London, E1 4NS, UK

²Trauma Unit, Royal London Hospital, Barts and The London NHS Trust, London, E1 1BB,
UK

barbaros@eeecs.qmul.ac.uk

Abstract. We describe a method of building a decision support system for clinicians deciding between interventions, using Bayesian Networks (BNs). Using a case study of the amputation of traumatically injured extremities, we explain why existing prognostic models used as decision aids have not been successful in practice. A central idea is the importance of modeling causal relationships, both so that the model conforms to the clinicians' way of reasoning and so that we can predict the probable effect of the available interventions. Since we cannot always depend on data from controlled trials, we depend instead on 'clinical knowledge' and it is therefore vital that this is elicited rigorously. We propose three stages of knowledge modeling covering the treatment process, the information generated by the process and the causal relationship. These stages lead to a causal Bayesian network, which is used to predict the patient outcome under different treatment options.

Keywords: Bayesian Networks, Causal Models, Clinical Decision Support

1 Introduction

How can a decision-support system assist a clinician deciding between several available treatments (or 'interventions') for a patient? We describe a method of building a decision support system applicable to this problem, based on the use of Bayesian Networks (BNs). Our focus here is on the prediction of the outcome for the patient, given the different treatment options, as if to answer a clinician asking "what is likely to happen to the patient if I do A or B?". Such a prediction is the first step needed to assist a decision maker; the further step from prediction to advice is not considered here.

We have developed the proposed method as part of a project to develop decision support for the treatment of traumatically injured (or 'mangled') extremities, where surgeons must decide whether or not to salvage or amputate the injured limb. We use this case study as a running example to illustrate each stage of the method.

The use of prognostic models in medicine is increasing [1]. Such models make predictions about the course of a disease from one or more predictors. The relationship between the predictors and the outcome does not always need to be causal [2]. On the other hand, when the need is to decide between possible interventions, a causal relationship between the intervention and the outcome is clearly necessary and this is a challenge when, as in our case study, we are depending on data gathered from past cases rather than from a controlled trial.

Randomised controlled trials (RCT) have been the primary way of identifying and measuring causal relations, since randomisation has the potential to reduce the effect of confounding variables. However, it is not straightforward to conduct RCTs for all questions of interest and the cost and time required for generalizable RCTs can be very high. The impracticality of RCTs is especially pertinent for an application such as the treatment of mangled extremity by amputation. Apart from the obvious practical and ethical issues, before an RCT is run some evidence of the potential benefits is needed and this must come from non-experimental sources.

Our proposal is to develop causal BNs based on a combination of expert medical knowledge and observational data. The knowledge is required to identify the causal relations and the data is used for determining the strengths of these relations. Knowledge is captured through a sequence of models describing the treatment process, the information available and a hierarchy of causal relationships.

The remainder of this paper is organised as follows: the case study about mangled extremity is first presented in Section 2, with Section 3 covering existing work on prognostic models and decision support for mangled extremity treatment. Section 4 presents the proposed method for building causal BNs. Conclusions and discussions are given in Section 5.

2 Case Study: Mangled Extremities

2.1 Treatment of Mangled Extremities

Clinicians often have to decide whether to amputate or salvage the extremity during mangled extremity treatment. This decision, with irreversible consequences for the patient, revolves around three possible adverse outcomes, which change in prominence as the treatment progresses.

1. **Death.** There is a risk to the patient's life from the injury to the limb. This risk depends on other injuries that may have been sustained at the same time. This risk is most prominent at the first stage of treatment.
2. **Limb tissue viability.** If the limb loses its blood supply for too long, its tissues becomes unviable and amputation becomes inevitable. The viability of the limb tissues is evaluated as the extent of the injury is accessed.
3. **Non-functional limb.** A salvaged limb may be more or less functional due to the anatomical problems such as loss of muscle compartments or transected nerves. For some patients a prosthetic limb may be preferable to a non-functional or

painful limb; this outcome becomes more prominent when it is clear that limb salvage is possible.

The clinician's concerns about these three treatment outcomes changes as the treatment progresses. The probabilities of the adverse outcomes are both positively and negatively related with each other so it may not be possible to find a decision that minimises all of them. For example, lengthy reconstruction surgeries can salvage patient's limb, but it can also put the patient's life in danger when the patient is physiologically unwell. In later stages of the treatment, following correction of initial physiology, infections of the damaged limb tissues may again threaten patient's life. Finally, the clinicians may decide to amputate the limb if it is not likely to be functional in the long run. Although the choice of treatment is the same, the underlying reasoning changes significantly through different stages of the treatment.

2.2 Experience of the Trauma Unit at the Barts and the London Hospital

The Royal London Hospital (RLH) is an internationally recognised leader in trauma care and trauma research. The trauma unit is the busiest in the United Kingdom treating over 2000 injured patients last year (2010), a quarter of whom were severely injured. The hospital is also the lead for a network of trauma hospitals, the London Trauma System, which provides specialist trauma care for the millions of people living in London and the South-East of England. This trauma system is believed to be the largest of its kind in the world. As a major trauma centre the hospital provides expedient access to the latest technology, treatments and expert trauma clinicians around the clock. Evidence has shown that people who suffer serious injuries need the highest quality specialist care to give them the best chances of survival and recovery.

The most common cause of injury seen at the Royal London Hospital is road traffic collisions followed by stabbings and falls from a height. Nearly half of the trauma patients have an injury to an extremity or the pelvic girdle, and 1% of these patients end up having lower limb amputations. A large multidiscipline team manages those with severe limb injuries. These devastating injuries carry a high mortality and morbidity in a predominantly young population. The multidiscipline approach ensures the best possible outcome for these patients.

2.3 Characteristics of this Decision Problem

We can summarise the characteristics of the limb amputation decision problem as follows:

- The treatment pathway is complex and the decision evolves with the treatment.
- Multiple outcomes need to be considered.
- The information relevant to the decision changes with time.

These characteristics suggest the need for analysis of the information available and modelling of the care pathway before a decision model can be developed.

3 Prognostic Models

3.1 Traditional Prognostic Models

Prognosis is the act of predicting the course of a disease or a medical condition. A prognostic model makes such predictions based on several independent predictors. Typically, the relation of the predictors to the model outcome is analysed by multivariate statistical models or similar approaches [3]. The accepted way of selecting predictors is to adjust the variables and check their effects on the outcome in observational data. If an adjustment of a variable is connected to the outcome with statistical significance, the variable can be called as an independent predictor. The danger is that correlation is confused with causation. For example, grey hair is an independent risk factor for heart disease, however, if two men of the same age but different hair colours are considered, grey hair does not probably increase the heart disease risk [2]. Therefore, the independent predictors are not necessarily causal factors; they are the factors that are correlated with causal factors according to the available data and selected variables. More extreme examples about variable selection can be seen in some scientific studies where electric-razors or owning refrigerators have been identified as risk factors for cancer [4]. Consequently, the independent predictors and their relations to outcome can be completely different between studies. Predictors with different sets of variables can be statistically accurate but high statistical accuracy of a model does not ensure its clinical acceptance [5] and there are now widely accepted arguments against the use of statistical significance tests and their associated p-values [6]. Clinicians demand models that have reasonable and understandable knowledge base aligned with latest clinical guidelines [7, 8].

On the other hand, there is an abundance of domain knowledge about the clinically relevant variables and their causal relations that can be integrated into model building. The main problems of traditional prognostic approaches can be overcome if domain knowledge is used.

3.2 Scoring Systems for Mangled Extremity Treatment

Multiple scoring systems have been developed as decision support models for mangled extremity treatment [9]. All of these models grade a patient's situation according to several injury-related variables. If a patient's score is above the model's threshold value, the model recommends an amputation. However, these scoring systems have not been widely accepted as a decision support tool by clinicians; we consider some reasons for this below.

Firstly, the scoring systems were developed based on observational data with low sample sizes. For example, MESS [10], which is a widely known scoring system, was developed with data on just 26 patients. Consequently, the high predictive results obtained by the authors were not repeated in later independent validation studies that have a higher number of participants (Table 1). Validation of the model was measured by sensitivity, which is the percentage of the amputated limbs that were also predicted to be amputated by the model, and by specificity, which is the percentage of the

salvaged limbs that were predicted as such by the model. Sensitivity and specificity results for the other scoring systems were similar as well. Bosse et al.'s multicentre prospective study [11] concluded that the predictive performance of the scoring systems was poor.

Table 1. Validation Studies for MESS

| <i>Validation Study</i> | <i>Participants</i> | <i>Sensitivity</i> | <i>Specificity</i> |
|-------------------------|---------------------|--------------------|--------------------|
| MESS's developers [10] | 26 | 1 | 1 |
| Robertson et al.[12] | 154 | 0.43 | 1 |
| Bonanni et al.[13] | 89 | 0.22 | 0.53 |
| Durham et al.[14] | 51 | 0.79 | 0.83 |
| Bosse et al.[11] | 556 | 0.46 | 0.91 |
| Korompilias et al.[15] | 63 | 0.87 | 0.71 |

Secondly, the output of scoring systems was the amputation decision itself. As a result, if there is a discrepancy between the model's recommendations and clinician's decisions, the model does not provide any useful decision support apart from implying that this outcome was the decision that was made in the model's training data. Thirdly, the scoring system's performance cannot be assessed in practice by sensitivity and specificity values since these measures represent the similarity between the models' recommendations and clinicians' decisions. A model can have 100% sensitivity and specificity but there is a possibility that both model and the compared clinicians were wrong.

3.3 Bayesian Networks

Bayesian networks (BNs) are probabilistic graphical models with multiple variables and relevant independence assumptions that are suitable for representing causality. All BNs, on the other hand, are not necessarily causal since the BNs can effectively represent non-causal probabilistic relations as well as the causal ones. BNs have been proposed for a wide range of medical applications [16] including prognosis [17] and prediction of the outcomes of different interventions [18].

Verduijn et al. [17] proposed a method for learning BNs specifically for prognosis from observational data. Their approach has several advantages compared to traditional prognostic models since it can represent the reasoning mechanism among intermediate variables. Moreover, in contrast to regression models the multiple stage nature of prognostic decisions can be implemented in BNs. Although Verduijn et al.'s prognostic BNs [17] are capable of learning more complex relations from observational data; those relations are still not necessarily causal so that making predictions about interventions is not possible. There are several methods for learning parts of causal relations from data [19] but these methods require extensive amount of data which may not be feasible for relatively uncommon medical conditions such as traumatic amputations.

Causal BNs [19] should have a clear relationship to the complex procedural, associational and hierarchical aspects of the clinical knowledge together with the

causal relations. Such knowledge is elicited and verified from multiple experts to minimise the biases. However, communicating through the model becomes more difficult with this additional complexity. Moreover, the risk of introducing a semantic mistake to the model increases.

Several knowledge modelling approaches have been proposed to overcome those difficulties in building BN structure. Nadkarni and Shenoy [20] outline a procedure which can be useful for building simpler causal BNs. Laskey and Mahoney [21] propose using systems engineering methods for building larger and more complex BNs. Object-oriented approaches have been proposed as well to assist the building of larger BNs [22, 23]. Laskey and Mahoney [24] propose using network fragments with object-oriented concepts to represent repeatable structures in the problem domain that are meaningful to the experts. Neil et al. [25] use repeatable structures that represent commonly encountered modelling tasks such as modelling the measurements in BNs. A more automated way of building BN with expert knowledge is proposed by Wiegerinck [26] in which constraints on the model are identified by the experts, and the model is modified by minimising a cost-function which shows the model's differences from those constraints. Although their method is primarily used for tuning model parameters, model structure can be modified as well with the help of the cost function and the constraints. Helsper and van der Gaag [27, 28] propose keeping detailed background knowledge for the BN in a separate ontology from which they gather initial BN alternatives. These alternatives are then modified and improved until one of them is satisfactory for the user. Additional expert knowledge, which is not stored in the ontology, could be necessary for these improvements. Moreover, it is not clear if the aims of the BN, relevant decisions and priorities could be analysed with the ontology. These issues should be clearly identified in the knowledge base for a complex multi-stage decision making problem like the mangled extremity treatment. In the following section, we will give some examples about the challenges of BN building which have not been fully solved by the previous knowledge modelling approaches, and introduce a method addressing them.

4 Knowledge Modelling for Causal Bayesian Networks

Since our proposal to use causal BNs depends on the elicitation of knowledge about causal relationships between variables, explicit knowledge modelling is central to our proposed method. In this section, we describe this knowledge modelling, illustrating it with examples from the case study of mangled extremities.

4.1 Method Overview

Our goal is to develop BN models to predict one or more outcome variables, depending on the values of other relevant factors and conditioned on the possible outcomes. The first imperative is therefore to have a clear understanding of all the variables in the model (i.e. *clarity test* [29]), so before constructing the BN we need to capture knowledge about the entities and attributes relevant to the domain. These

entities may relate to different stages of the treatment process and some attributes may have changing values. A complete understanding of the data therefore depends on knowledge of the treatment process. Moreover, the predictions needed for decision support may change through the treatment. A model of this process is therefore our starting point.

4.2 Modelling the Treatment Process

Decisions about clinical interventions are usually done in iterative stages until the patient is treated. After making an intervention, clinicians observe the results of the intervention, re-evaluate treatment risks, and select a treatment alternative [30]. Activity diagrams (Fig. 1) can be used to identify the decisions that are important for the clinical problem, and the priorities of these decisions throughout the treatment.

The changing decision priorities in mangled extremities are illustrated by an example about a patient treated by surgeons at RLH following a motor-cycle accident that resulted in severe leg injury and serious bleeding. When the patient arrives at the hospital, his physiology is in a dangerous condition due to bleeding but his limb appears to be anatomically salvageable. A causal BN used for decision-support at this stage will access the physiology-related risk of death, considering the options of a reconstruction operation, and the possibility of salvaging the limb later. Consequently variables of the model will be mainly about physiology, bleeding and limb injury.

The risk of death related to physiology may decrease if the patient is resuscitated for a few days. However, other risks to the patient's life may develop in the following days. These include infections and renal failure resulting from dead or dying tissues. The causal BN used at this stage will still provide decision support about the risk of death and possibility of limb salvage but its predictions will be based these developing pathologies. If the risk of death related to limb injury is also low, the clinicians will evaluate the possibility of anatomical salvage and future functioning of the limb. The causal BN for this stage will be more focused on structure of the injured limb rather than mechanisms related to death.

Modelling such differences between decision making stages could be complex, especially if there are multiple decisions with various priorities and interrelated outcomes like the mangled extremity treatment example. The activity diagrams provide a clear and understandable map of the main decisions making stages and relevant interventions (shown by diamonds and rectangles respectively in Fig. 1). Most of the medical experts are familiar with the format of activity diagrams as many clinical guidelines are published in similar ways. Therefore, it is also a convenient communication medium between the domain expert and the knowledge engineer.

Not all the information may be needed in the BN to predict the outcomes of interest at each stage. The main outcomes and relevant variables for each stage can be identified by using the activity diagram. For example, when we focus on the first decision making stage about the patient's physiology with the domain experts, the main outcome for this stage is identified as the risk of death, and the relevant clinical variables are identified as bleeding, shock, and coagulation. The main causal relations between the outcome and the clinical variables are identified as: the patient enters a

lethal state of circulatory shock as a result of bleeding, shock may impair the body's coagulation ability worsening the future course of bleeding and shock status.

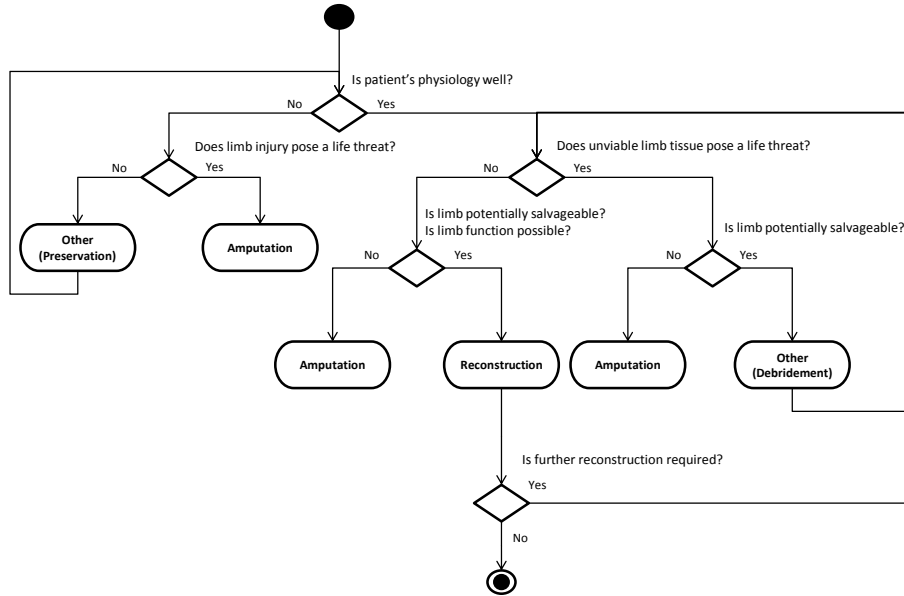


Fig. 1. Activity Diagram for Mangled Extremity Decision Making

4.3 Modelling Information Arising From Treatment

The variables used in the BN must be clearly defined, corresponding to an attribute of a defined entity, at a given stage of treatment. Information models that represent the knowledge about relevant entities and their attributes can guide the selection of variables in the BN. Moreover, multiplicity about these variables must be clarified as well. In our case study, a patient may have an amputation in each of their two limbs. Moreover, the same limb could be sequentially amputated at progressively higher levels. For example, there are records for 53 patients, 73 limbs and 83 amputation operations in the data from RLH about lower limb amputations.

The information model can be used with the activity diagram to identify the variables relevant to each decision making stage. For example, we identified main variables and causal relations for the first decision making stage in Section 4.2 but some of those variables (e.g. shock) are unobservable so that their states must be estimated by other observable attributes (e.g. systolic blood pressure (SBP), heart rate (HR), respiratory rate (RR)). The information model can be used to identify such attributes related to main entities (Fig. 2).

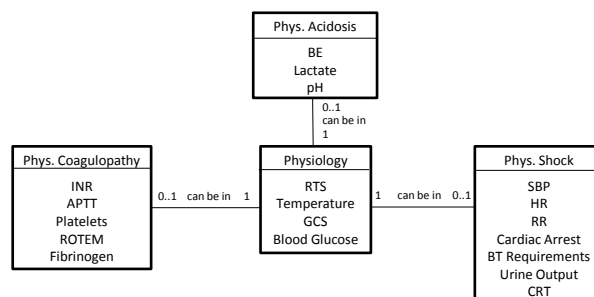


Fig. 2. Fragment of the Information Model about Physiology

Many of the unobservable variables about physiology and their estimators are continuously changing variables but the values for most of these are not measured continuously. For example, multiple blood tests are used to estimate the changes in coagulopathy. Moreover, the causal BN models are used in discrete time stages therefore the relations between the variables with multiple measurements and their representations in the causal BN must be clarified. The class diagrams can be used for illuminating those relations. In the class diagram about mangled extremity treatment (Fig. 3), the model assumes that a patient can have multiple interventions, and the patient's physiology status can change between these interventions. Therefore, the instantiations about a patient's physiology in the causal BN shows the state in each intervention. On the other hand, variables about a patient's past medical history (PMH) or injury are static (Fig. 3) thus they have single fixed values in the causal BN.

4.4 Model Causal Relationships at Different Knowledge Levels

While clinicians usually express their reasoning in small and compact statements, these statements are actually based on series of cause-effect deductions from more complex structures. Methods for representing multiple levels of clinical knowledge have been developed [31]. The causal BNs with less detail abstract the detailed information about a part of a clinical problem. These models can show the main causal relations with fewer variables which is suitable for communication with the experts about the overall model structure. More detailed causal BNs can show more complex relations that could be used for making inferences about detailed mechanisms if there is available data (for example, from a variety of laboratory tests). These models are aligned by the less detailed models through focal nodes. Focal nodes are anchors for the different knowledge levels that describe the same concept and share the same name [31].

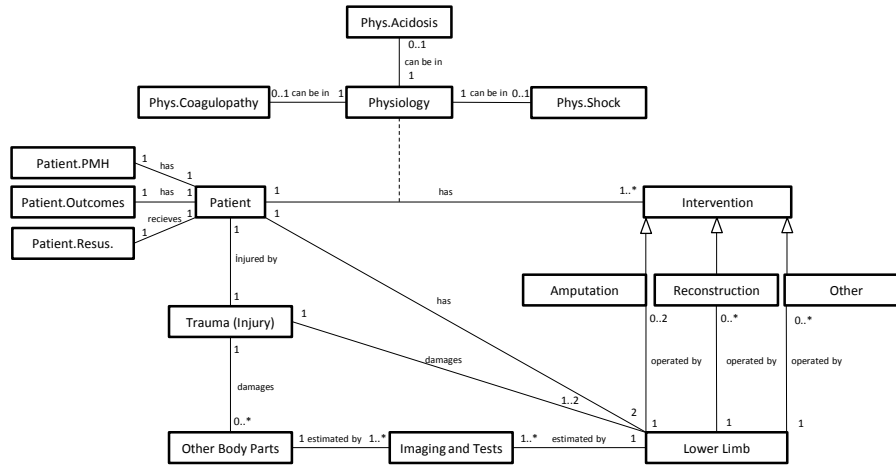


Fig. 3. Class diagram of entities related to mangled extremity treatment

An example of causal BNs with different detail levels is shown in a fragment of the mangled extremity model (Fig. 4). These causal BNs model a part of the physiology related risk of death which is crucial in early stages of the treatment (Fig. 1). The outcomes (death) and main variables (bleeding, shock, coagulopathy) for the model were identified in Section 4.2 with the help of the activity diagram. The feedback relationship between coagulopathy and future course of bleeding has not been represented as a dynamic BN in this illustration for simplicity. The least detailed causal BN shows the overall causal relations between bleeding, circulatory shock, coagulopathy, the risk of death and possible interventions i.e. amputation or rapid surgery. Although this model represents the overall causal relationships, it does not show the two intermediate (temperature, acidosis) variables between shock and coagulopathy. A more detailed version of the causal BN can be built by adding these relations as well as the estimators for the unobservable shock variable (RR, HR, SBP, capillary refill time (CRT), urine output, Glasgow coma scale (GCS)) which were identified by the information model in Section 4.3 (Fig. 2). This model could bring more explanatory predictions due to additional causal mechanisms. The relation between shock and its seven estimators can also be explained in a more detailed way. For example, urine output that is used for estimating shock is caused by perfusion in the kidneys. The increase in respiratory rate is caused by lack of O₂ delivery to the tissues as a result of low perfusion. Therefore, knowledge detail in the model can be increased by modelling shock through these relations. However, estimating values about the perfusion in different body parts could be more difficult for the user than estimating a value for shock only. The nodes that are not modelled in different levels of details, such as bleeding or coagulopathy node in our example can be used as focal points to align the models and keep the overall causal relations consistent between different detail levels.

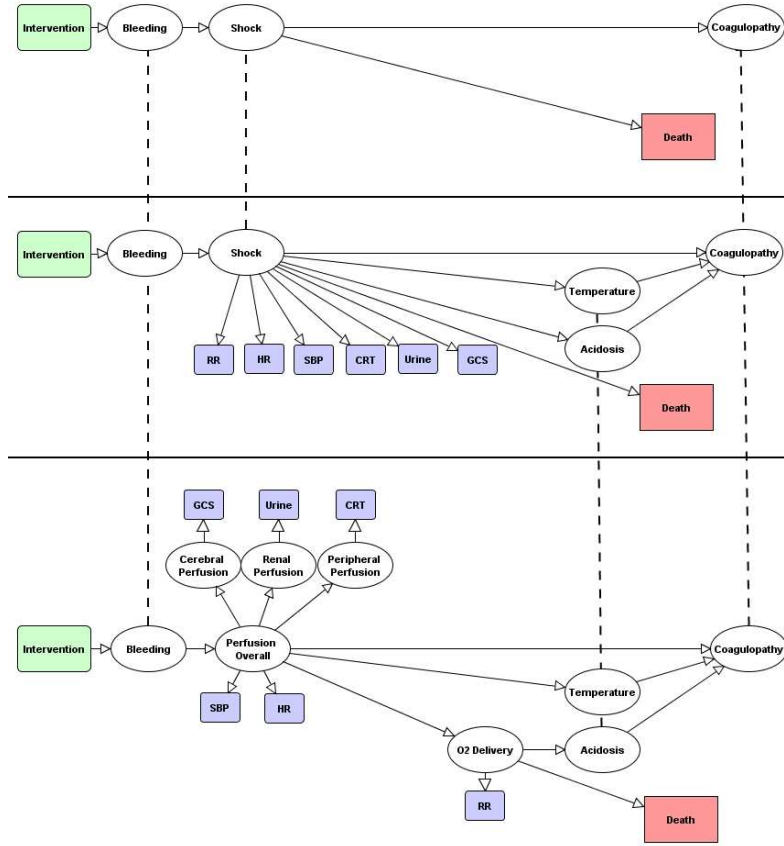


Fig. 4. Causal BN with multiple levels of detail about physiology related risk of death

Any of the three models in the example can be used depending on the available information about the variables, predictive performance of the model and preference of the user. The multi-level causal BN makes it possible to keep a consistent and understandable knowledge-base for the model regardless of the modelling preferences. This could be useful in improving the model's clinical acceptance since a clear and reliable knowledge-base is one of the main demands from the prognostic models [7, 8].

4.5 Modelling Dynamic Variables

Many continuous clinical variables are estimated by multiple discrete measurements such as blood tests. The multiplicity relations between these variables and their measurements can be identified by the information models shown in Section 4.3. On the other hand, modelling the effects of continuously changing variables in the BN still remains an issue. One well-known solution for this issue is to instantiate the complete model structure over multiple time slices. However, this approach could be

computationally infeasible if there are numerous time stages and large model structures.

One approach for modelling continuously changing variables in the BN could be to use trend variables that summarise the variations of several previous instantiations of the related variable. In clinical practice, the trends of historical measurements for some clinical factors are used to make predictions about patient outcomes. For example, a patient’s response to resuscitation, which can be analysed by trends of several diagnostic values about shock, is an important factor for predicting the patient’s survival. This can be modelled in BN by adding a trend variable that summarises the variations in the previous states of shock (Fig 5.).

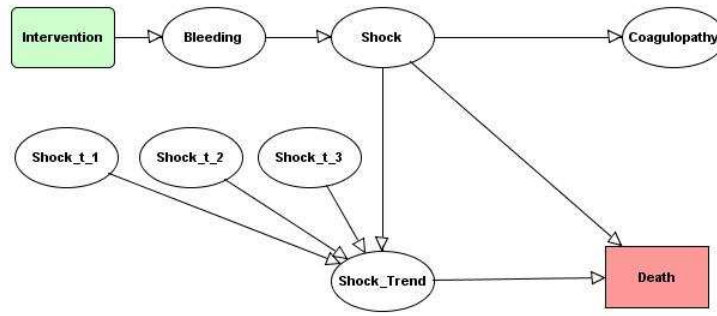


Fig. 5. Fragment of the causal BN with a trend variable

5 Conclusion

In this study, we have proposed a method for building causal BNs, where causal relationships are elicited from clinical knowledge. The method involves three stages of knowledge modelling, using:

- activity diagrams to model the decision points and procedural relations
- class diagrams to model the multiplicity relations between the variables
- multi-level causal diagrams to represent a hierarchical of causal relationships.

This method aids the knowledge-elicitation with experts by providing understandable intermediate models and decreases the risk of having semantic mistakes in the final BN model. The study for developing the method is still in progress. This paper shows our first attempts for providing guideline for some common modelling problems seen in building causal BNs. More structured method for building complete causal BNs are being researched. For next steps, we plan to formalise the models within a common framework, allowing more automated approaches for building the final causal BNs. The outcomes of the causal BN are posterior probability distributions about the treatment risks in a variety of situations. Although these posterior distributions can provide useful information for the decision maker, we plan to analyse these distributions’ relations to decision making and prepare clinical guidelines that are more helpful and efficient for the decision maker.

Acknowledgements

We are grateful for the contribution for Mr Nigel Tai, FRCS, Consultant Trauma and Vascular Surgeon at the Barts and the London NHS Trust to the work described in this paper.

References

1. Moons, K., Royston, P., Vergouwe, Y., Grobbee, D., Altman, D.: Prognosis and prognostic research: what, why, and how? *BMJ* 338, 1317--1320 (2009)
2. Brotman, D.J., Walker, E., Lauer, M.S., O'Brien, R.G.: In Search of Fewer Independent Risk Factors. *Arch. Intern. Med.* 165, 138--145 (2005)
3. Abu-Hanna, A., Lucas, P.J.F.: Prognostic Models in Medicine. *Method. Inform. Med.* 40, 1--5. (2001)
4. Jenks, S., Volkers, N.: Razors and refrigerators and reindeers – oh my. *J. Natl. Cancer. Inst.* 84, 1863 (1992)
5. Moons, K.G.M., Altman, D.G., Vergouwe, Y., Royston, P.: Prognosis and prognostic research: application and impact of prognostic models in clinical medicine. *BMJ* 338, 1487--1490 (2009)
6. Ziliack, S.T., McCloskey, D.N.: The cult of statistical significance : How the standard error costs us job, justice, and lives. University of Michigan Press, Ann Arbor (2008)
7. Wyatt, C.J., Altman, D.G.: Commentary: prognostic models: clinically useful or quickly forgotten? *BMJ* 311, 1539--1541 (1995)
8. Akhthar, J.I., Forse, R.A.: Prognostic models: Are these models health fortune-telling tools? *Crit. Care. Med.* 38(7), 1605--1606 (2010)
9. Rush, R.M. Jr., Beekley, A.C., Puttler, E.G., Kjorstad, R.J.: The Mangled Extremity. *Curr. Probl. Surg.* 46(11), 851--926 (2009)
10. Johansen, K., Daines, M., Howey, T., Helfet, D., Hansen, S.T.Jr.: Objective criteria accurately predict amputation following lower extremity trauma. *J. Trauma.* 30, 568--572 (1990)
11. Bosse, M.J., MacKenzie, E.J., Kellam, J.F., Burgess, A.R., Webb, L.X., Swiontkowski, M.F., Sanders, R.W., Jones, A.L., McAndrew, M.P., Patterson, B.M., McCarthy, M.L., Cyril, J.K.: A Prospective Evaluation of the Clinical Utility of the Lower-Extremity Injury-Severity Scores. *J. Bone. Joint. Surg.* 83-A(1), 3--14 (2001)
12. Robertson, P.A.: Prediction of Amputation after Severe Lower Limb Trauma. *J. Bone. Joint. Surg.* 73-B(5), 816--818. (1991)
13. Bonanni, F., Rhodes, M., Lucke, J.F.: The Futility of Predictive Scoring of Mangled Lower Extremities. *J. Trauma.* 34, 99--104 (1993)
14. Durham, R.M., Mistry, B.M., Mazuski, J.E., Shapiro, M., Jacobs, D.: Outcome and Utility of Scoring Systems in the Management of the Mangled Extremity. *Am. J. Surg.* 172, 569--574 (1996)

15. Korompilias, A.V., Beris, A.E., Lykissas, M.G., Vekris, M.D., Kontogeorgakos, VA, Soucacos, PN: The mangled extremity and attempt for limb salvage. *J. Orthop. Surg. Res.* 4(4). (2009)
16. Lucas, P.J.F, van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. *Artif. Intell. Med.* 30(3), 201--214 (2004)
17. Verduijn, M., Peek, N., Rosseel, P.M.J., de Jonge, E., de Mol, B.A.J.M.: Prognostic Bayesian Networks I: Rationale, learning procedure and clinical use. *J. Biomed. Inform.* 40(6), 609--618 (2007)
18. Fenton, N., Neil, M.: Comparing risks of alternative medical diagnosis using Bayesian arguments. *J. Biomed. Inform.* 43, 485--495 (2010)
19. Pearl, J.: *Causality*. Cambridge University Press, Cambridge. (2000)
20. Nadkarni, S., Shenoy, P.P: A Bayesian network approach to making inferences in causal maps. *Eur. J. Oper. Res.* 128, 479--498 (2001)
21. Laskey, K., Mahoney, S.: Network engineering for agile belief networks. *IEEE T. Knowl. Data. En.* 12, 487--498 (2000)
22. Koller, D., Pfeffer, A.: Object-Oriented Bayesian Networks. In: *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pp. 302--313. (1997)
23. Bangsø, O., Willemin, P.H.: Top-down Construction and Repetitive Structures Representation in Bayesian Networks. In: *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference*, pp. 282--286. (2000)
24. Laskey, K., Mahoney, S.: Network Fragments: Representing Knowledge for Constructing Probabilistic Models. In: *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pp. 334--341.
25. Neil, M., Fenton, N., Nielsen, L.: Building large-scale Bayesian networks. *Knowl. Eng. Rev.* 15(3), 287--284. (2000)
26. Wiegerinck, W.: Modeling Bayesian Networks by Learning from Experts. In: *Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2005)*, pp. 305--310. (2005)
27. Helsper, E.M., van der Gaag, L.C.: Building Bayesian networks through ontologies. In: *15th European Conference on Artificial Intelligence*, pp. 680--684. (2002)
28. Helsper, E.M., van der Gaag, L.C.: Ontologies for probabilistic networks: a case study in the oesophageal-cancer domain. *Knowl. Eng. Rev.* 22(1), 67--86. (2007)
29. Kjørulff, U.B., Madsen, A.L.: *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer, New York. (2008)
30. Sox, H.C., Marshal, A.B., Michael, C.H., Keith, I.M.: *Medical Decision Making*. ACP, Philadelphia. (1988)
31. Patil, R.S., Szolovitz, P., Schwartz, W.B.: Causal Understanding of Patient Illness in Medical Diagnosis. In: *Seventh International Joint Conference on Artificial Intelligence*, pp. 893--899. (1981)

Cost-effectiveness analysis with influence diagrams

M. Arias and F. J. Díez

UNED, Madrid, Spain

Abstract. Cost-effectiveness analysis (CEA) is used more and more frequently in medicine to determine whether the health benefits of an intervention outweighs its economic cost. In this paper we present two algorithms for performing CEA with influence diagrams; one of them is based on variable elimination and the other on arc reversal. Using the former, we have performed CEA on two influence diagrams whose equivalent decision trees contain thousands of leaves.

Keywords: Cost-effectiveness analysis, net monetary benefit, influence diagrams.

1 Introduction

In medicine, one of the methods of assessing whether the health benefits of an intervention outweighs its economic cost is cost-effectiveness analysis (CEA) [4, 5]. In this context, the *net monetary benefit* [16] of an intervention I_i is

$$NMB_{I_i}(\lambda) = \lambda \cdot e_i - c_i, \quad (1)$$

where e_i is its effectiveness and c_i its cost. The parameter λ is used to convert the effectiveness into a monetary scale. It takes values on the set of positive real numbers, i.e., in the interval $(0, +\infty)$. It is measured in effectiveness units divided by cost units; for example, in dollars per death avoided or euros per quality adjusted life-year (QALY [19]). It is sometimes called *willingness to pay*, *cost-effectiveness threshold* or *ceiling ratio*, because it indicates how much money a decision maker accepts to pay to obtain a certain “amount” of health benefit.

When the consequences of an intervention are not deterministic, it is necessary to apply a model that takes into account the probability of each outcome. The most usual tool for modeling decision problems with uncertainty are decision trees [13]. In a previous paper [1] we have presented a method of performing CEAs on decision trees with an arbitrary number of decision nodes.

The main drawbacks of decision trees is that the size of a tree grows exponentially with the number of variables, they cannot represent conditional independencies, and they require a preprocessing of the probabilities [2, 6]; for instance, medical diagnosis problems are usually stated in terms of direct probabilities, namely the prevalence of the diseases and the sensitivity and specificity of the tests, while the tree is built with the inverse probabilities, i.e., the positive and

negative predictive values of the tests. Even in cases with only a few chance variables, this preprocessing of probabilities is difficult, if not impossible.

An alternative representation language for decision making are influence diagrams (IDs) [6]. They have the advantages of being very compact, representing conditional independencies, and using direct probabilities. However, the only algorithm that can perform CEA with IDs is that proposed by Nielsen et al. [11], which is very difficult to apply in practice for the reasons discussed below. In this paper, we present two efficient algorithms for CEA with IDs that have allowed us to solve medical problems that were impossible to address with the techniques available so far. One of them is based on the variable elimination algorithm [7]; the other, on the arc reversal algorithm [12, 14].

The rest of this paper is structured as follows: Section 2 reviews the basic concepts of CEA and IDs. The new algorithms are presented in Section 3. Section 4 shows an example, Section 5 discusses some related work, and Section 6 contains the conclusions.

2 Background

2.1 Cost-effectiveness analysis (CEA)

Deterministic CEA Cost-effectiveness analysis (CEA) consists of finding the intervention that maximizes the net benefit for each value of λ (cf. Eq. 1). When we have a set of interventions such that the cost and effectiveness of each one are known with certainty, we can perform a deterministic cost-effectiveness analysis, which returns the optimal intervention for each interval of the possible values of λ . The standard algorithm for this analysis consists of eliminating the interventions dominated by another intervention (simple dominance), then eliminating the interventions dominated by a pair of other interventions (extended dominance), and finally computing the incremental cost-effectiveness ratios—see [18] or any book on medical decision analysis. This algorithm and an alternative method for deterministic CEA can be found in [1].

CEA with decision trees Sometimes we do not know explicitly the cost and effectiveness of each intervention, but we do know that each one may lead to different outcomes with different probabilities, which may in turn cause other outcomes, each having a known cost and effectiveness. In this case, we can build a decision tree such that each node, instead of representing a single utility, represents the cost and effectiveness of the corresponding scenario.

If the only decision node of the tree is the root, the tree can be evaluated by a modified version of the roll-back algorithm that computes the cost and effectiveness of each node separately and then performs a deterministic CEA at the root node [4, 15]. If the tree contains embedded nodes, this method cannot be applied because the evaluation of a decision node does not return a cost-effectiveness pair. However such a tree can be evaluated with the algorithm proposed in [1].

2.2 Influence diagrams

Basic properties of IDs An ID is a probabilistic graphical model that consists of three disjoint sets of nodes: decision nodes \mathbf{V}_D , chance nodes \mathbf{V}_C , and utility nodes \mathbf{V}_U . Chance nodes represent events that are not under the direct control of the decision maker. Decision nodes correspond to actions under the direct control of the decision maker. Given that each chance or decision node represents a variable, we will use indifferently the terms variable and node. Standard IDs require that there is a total ordering of the decisions, which indicates the order in which the decisions are made.

In this section, we assume that the ID contains only one utility node. However, the algorithms presented here can be easily extended to IDs in which the global utility is the sum of the values represented by the utility nodes [7] and to IDs containing *super-value* nodes [10, 17]. (A utility node is said to be *super-value* if its parents are other utility nodes.)

The meaning of an arc in an ID depends on the type of nodes that it links. An arc $X \rightarrow C$ where C is a chance node denotes a probabilistic dependence of C on X ; in practice, it usually means that X is a cause of C . An arc from a decision D_i to a decision D_j means that D_i is made before D_j . An arc from a chance node C to a decision node D_j means that the value of variable C is known when making decision D_j . Standard IDs assume the *non-forgetting hypothesis*, which means that a variable C known for a decision D_j is also known for any posterior decision D_k , even if there is not an explicit link $C \rightarrow D_k$ in the graph. An arc from a variable X to the utility node means that the utility depends on X .

A *potential* is a real-valued function over a domain of finite variables. The quantitative information that defines an ID is given by assigning to each chance node C a conditional probability potential $P(c|pa(C))$ for each configuration of its parents, $pa(C)$ and assigning to the utility node a potential $U(pa(U))$ that maps each configuration of its parents onto a real number.

The total ordering of the decisions $\{D_1, \dots, D_n\}$ induces a partition of the chance variables $\{\mathbf{C}_0, \mathbf{C}_1, \dots, \mathbf{C}_n\}$, where \mathbf{C}_i is the set of variables unknown for D_i and known for D_{i+1} . The set of variables known to the decision maker when deciding on D_i is called the *informational predecessors* of D_i and denoted by $IPred(D_i)$. Consequently,

$$IPred(D_i) = \mathbf{C}_0 \cup \{D_1\} \cup \mathbf{C}_1 \cup \dots \cup \{D_{i-1}\} \cup \mathbf{C}_{i-1} \quad (2)$$

$$= IPred(D_{i-1}) \cup \{D_{i-1}\} \cup \mathbf{C}_{i-1} . \quad (3)$$

The *maximum expected utility* (MEU) of an ID whose chance and decision variables are all discrete is

$$MEU = \sum_{\mathbf{c}_0} \max_{d_1} \sum_{\mathbf{c}_1} \dots \sum_{\mathbf{c}_{n-1}} \max_{d_n} \sum_{\mathbf{c}_n} \prod_{C \in \mathbf{V}_C} P(c|pa(C)) \cdot U(pa(U)) . \quad (4)$$

A *policy* δ_{D_i} is a function that maps each configuration of informational predecessors of D_i onto a value d_i of D_i . The *optimal policy* δ_{D_i} for decision D_i

is given by the following equation (in the case of a tie, any of the values of D_i that maximize that expression can be chosen arbitrarily):

$$\begin{aligned} & \delta_{D_i}(IPred(D_i)) \\ &= \arg \max_{d_i} \sum_{\mathbf{c}_i} \max_{d_{i+1}} \dots \sum_{\mathbf{c}_{n-1}} \max_{d_n} \sum_{\mathbf{c}_n} \prod_{C \in \mathbf{V}_C} P(c|pa(C)) \cdot U(pa(U)) . \end{aligned} \quad (5)$$

Algorithm 1: Variable elimination with division of potentials

Input: An influence diagram.
Result: The expected utility and the optimal policy for each decision.
// Initialize the list of probability potentials
1 $list \leftarrow \{P(c|pa(C)) \mid C \in \mathbf{V}_C\};$
2 **for** $i \leftarrow n$ **to** 0 **do**
 // eliminate the variables in \mathbf{C}_i
3 **foreach** *variable* $C \in \mathbf{C}_i$ **do**
4 take out from $list$ all the potentials that depend on C ;
5 $\psi \leftarrow$ product of those potentials;
6 $\psi_{ind} \leftarrow \sum_c \psi$;
7 $\psi_{cond} \leftarrow \psi / \psi_{ind}$;
8 add ψ_{ind} to the $list$;
9 $U \leftarrow \sum_c \psi_{cond} \cdot U$;
10 **if** $i > 0$ **then**
 // eliminate the decision D_i
11 $U \leftarrow \max_{d_i} U$;
12 $\delta_{D_i}(IPred(D_i)) = \arg \max_{d_i \in D_i} U$;
13 **foreach** *potential* $P(c|pa(C)) \in list$ **do**
14 **if** *this potential depends on D_i* **then**
15 project this potential onto the remaining variables
16 in the $list$;
17 replace $P(c|pa(C))$ with its projection;
18 **return** U

Variable elimination The direct application of the above expression leads to a computational cost that grows exponentially with the number of variables in the ID. A more efficient approach consists of eliminating the variables one by one, in an order compatible with the above equations, i.e., eliminating first the variables that appear on right-most operators of summation and maximization in Equation 4, as indicated by the Algorithm 1. The details and the justification of this algorithm can be found in [7, 10].

Please note that the potential ψ , which is the product of all the potentials that depend on C (line 5), is factored into two potentials: ψ_{ind} , where “ind”

stands for “independent of C ”, and ψ_{cond} , where c stands for “conditional probability” because this potential represents the conditional probability of C given the variables that have not been eliminated yet: $\psi_{\text{cond}} = P(c|\mathbf{v}_C)$. The meaning of this probability can be better understood if we consider that for each influence diagram there is an equivalent symmetric decision tree. The variables eliminated before C when evaluating the ID are placed on the right side of C in the decision tree, and those eliminated after C are placed on the left side. Each configuration \mathbf{v}_C represents a path from the root node to a node C in the tree, and the probability $P(c|\mathbf{v}_C)$ computed in the evaluation of the ID is the probability of the branches outgoing from that node in the tree.

Then the algorithm multiplies the potentials $\psi_{\text{cond}} = P(c|\mathbf{v}_C)$ and U , and sums out the variable C (line 9). This is equivalent to evaluating all the C nodes in the tree by computing the average of the utilities of their branches, using $P(c|\mathbf{v}_C)$ as the weights. The elimination of variable D_i by maximizing U (line 11) is equivalent to evaluating all the D_i nodes in the tree.

When all the chance and decision variables have been eliminated, the potential U contains only one real number, that is the expected utility of the ID (line 18).

Arc reversal An alternative method for evaluating IDs is the arc reversal algorithm proposed by Olmsted [12] (see also [14]). This algorithm consists of four basic operations:

1. **Barren node removal.** A node is said to be *barren* if it has no children. Barren nodes can be removed from the ID without performing any additional operation.
2. **Chance node removal.** A chance node Y whose only child is the utility node U can be removed from the ID by drawing links from each of its parents to U ; if \mathbf{X} is the set of parents of Y , $P_{\text{old}}(y|\mathbf{x})$ is the conditional probability of Y , \mathbf{Z} is the set of parents of U , and $U_{\text{old}}(\mathbf{z})$ is the utility potential before eliminating Y , the new utility potential after eliminating Y is

$$U_{\text{new}}(\mathbf{v}) = \sum_y P_{\text{old}}(y|\mathbf{x}) \cdot U_{\text{old}}(\mathbf{z}) , \quad (6)$$

where $\mathbf{V} = \mathbf{X} \cup (\mathbf{Z} \setminus \{Y\})$.

3. **Decision node removal.** A decision node D whose only child is the utility node U can be removed from the ID by drawing links from each of its parents to U ; if \mathbf{X} is the set of parents of D , \mathbf{Z} is the set of parents of U , and $U_{\text{old}}(\mathbf{z})$ is the utility potential before eliminating D , the new utility potential after eliminating D is

$$U_{\text{new}}(\mathbf{v}) = \max_d U_{\text{old}}(\mathbf{z}) , \quad (7)$$

where $\mathbf{V} = \mathbf{X} \cup (\mathbf{Z} \setminus \{D\})$.

4. **Arc reversal.** A link $X \rightarrow Y$ in the ID such that there is no other directed from X to Y in the graph can be reversed, i.e., replaced by $Y \rightarrow X$, by performing the following additional operations: (1) all the parents of X become

parents of Y , and vice versa, (2) the new probability of Y is

$$P_{\text{new}}(y|\mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_x P_{\text{old}}(x|\mathbf{a}, \mathbf{b}) \cdot P_{\text{old}}(y|\mathbf{b}, \mathbf{c}), \quad (8)$$

and the new probability of X is

$$P_{\text{new}}(x|y, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \sum_y \frac{P_{\text{old}}(x|\mathbf{a}, \mathbf{b}) \cdot P_{\text{old}}(y|\mathbf{b}, \mathbf{c})}{P_{\text{new}}(y|\mathbf{a}, \mathbf{b}, \mathbf{c})}, \quad (9)$$

where $\mathbf{A} = Pa(X) \setminus Pa(Y)$, $\mathbf{B} = Pa(X) \cap Pa(Y)$, and $\mathbf{C} = Pa(Y) \setminus \{Pa(X) \cup X\}$.

Each of these operations transforms an ID into an equivalent ID having the same optimal policies and the same expected utility (except for the decisions removed, obviously). The Theorem 4 in [14] states that if the ID does not contain any node that can be removed directly, there exists a sequence of arc reversals leading to an equivalent ID in which at least one node can be removed. The elimination of a decision node D gives the optimal policy for D . The utility potential remaining after eliminating all the chance and decision nodes contains a single real number (a scalar), that is the expected utility of the ID. The algorithm always terminates because the number of nodes in the ID is finite.

3 Cost-effectiveness analysis with IDs

3.1 Cost-effectiveness partitions

As mentioned above, cost-effectiveness analysis consists of finding an intervention that maximizes the net benefit for each value of λ ; in practice, it consists of finding the intervals for which an intervention is more beneficial than the others. We formalize this idea by introducing the concept of *cost-effectiveness partition*, CEP.

Definition 1. A cost-effectiveness partition (CEP) of n intervals is a tuple $Q = (\Theta_Q, C_Q, E_Q, I_Q)$, where:

- $\Theta_Q = \{\theta_1, \dots, \theta_{n-1}\}$ is a set of $n - 1$ values (thresholds), such that $0 < \theta_1 < \dots < \theta_{n-1}$,
- $C_Q = \{c_0, \dots, c_{n-1}\}$ is a set of n values (costs),
- $E_Q = \{e_0, \dots, e_{n-1}\}$ is a set of n effectiveness values, and
- $I_Q = \{I_0, \dots, I_{n-1}\}$ is a set of n interventions.

For the sake of simplifying the exposition, we define $\theta_0 = 0$ and $\theta_n = +\infty$ for every CEP.

Alternatively, a CEP can be denoted by a set of n 4-tuples of the form (interval, cost, effectiveness, intervention),

$$Q = \{((0, \theta_1), c_0, e_0, I_0), \\ ((\theta_1, \theta_2), c_1, e_1, I_1), \\ \dots, \\ ((\theta_{n-1}, +\infty), c_{n-1}, e_{n-1}, I_{n-1})\},$$

which means that when λ is in the interval (θ_i, θ_{i+1}) the most beneficial intervention is I_i , which has a cost c_i and an effectiveness c_i . When $\lambda = \theta_{i+1}$, there is a tie between I_i and I_{i+1} .

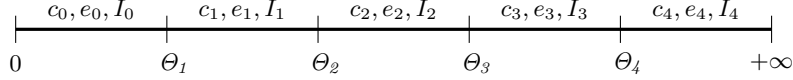


Fig. 1. Cost-effectiveness partition (CEP).

The functions $cost_Q(\lambda)$, $eff_Q(\lambda)$, $NMB_Q(\lambda)$, and $interv_Q(\lambda)$ return the cost, the effectiveness, the NMB, and the optimal intervention for λ according to the CEP Q ; see [1] for a mathematical definition of these functions.

Combination of cost-effectiveness partitions In this subsection we generalize the average and maximization operations from single utilities to CEPs.

Definition 2 (Weighted average). *Given a set of m CEPs $\{Q_1, \dots, Q_m\}$, a chance variable X whose domain is $\{x_1, \dots, x_m\}$, and a probability distribution for X , $P(x_j)$, we say that a CEP Q is a weighted average of the CEPs if*

$$\forall \lambda, \quad cost_Q(\lambda) = \sum_{j=1}^m P(x_j) \cdot cost_{Q_j}(\lambda) \quad (10)$$

and

$$\forall \lambda, \quad eff_Q(\lambda) = \sum_{j=1}^m P(x_j) \cdot eff_{Q_j}(\lambda). \quad (11)$$

A straightforward consequence of this definition is that, because of Equation 1,

$$\forall \lambda, \quad NMB_Q(\lambda) = \sum_{j=1}^m P(x_j) \cdot NMB_{Q_j}(\lambda). \quad (12)$$

These three equalities mean that for every value of λ , the cost, effectiveness, and NMB of the weighted average partition Q are the same as if we had performed a

weighted average for the values of cost and effectiveness of the Q_j 's. The partition Q can be efficiently computed by the Algorithm 2.

The intervention composed at the fifth line of the algorithm means: “if the chance variable X takes on the value x_j , then follow the policy indicated by the corresponding branch ($X = x_i$) of the tree”.

Algorithm 2: Weighted average of CEPs

Input: A set of m CEPs $\{Q_1, \dots, Q_m\}$, with $Q_j = (\Theta_j, C_j, E_j, I_j)$,
a chance variable X whose domain is $\{x_1, \dots, x_m\}$, and
a probability distribution for X , $P(x_j)$.

Result: A new CEP $Q = (\Theta, C, E, I)$.

```

1  $\Theta \leftarrow \bigcup_{j=1}^m \Theta_j$ 
2  $n \leftarrow \text{card}(\Theta)$ 
3 for  $i \leftarrow 1$  to  $n$  do
4    $c_i \leftarrow \sum_{j=1}^m P(x_j) \cdot \text{cost}_{Q_j}(\theta_i)$ 
5    $e_i \leftarrow \sum_{j=1}^m P(x_j) \cdot \text{eff}_{Q_j}(\theta_i)$ 
6    $I_i \leftarrow$  “If  $X = x_1$ , then  $\text{interv}_{Q_1}(\theta_i)$ ; if  $X = x_2$ , then  $\text{interv}_{Q_2}(\theta_i)$ ; etc.”
```

Definition 3 (Optimal partition). *Given a set of m CEPs $\{Q_1, \dots, Q_m\}$ and a decision D whose domain is $\{d_1, \dots, d_m\}$, a CEP Q is optimal if*

$$\forall \lambda, \exists j, \text{NMB}_{\text{interv}_{Q_j}(\lambda)}(\lambda) = \max_{j'} \text{NMB}_{\text{interv}_{Q_{j'}}(\lambda)}(\lambda), \quad (13)$$

$$\text{interv}_Q(\lambda) = \text{“choose option } d_j; \text{ then apply } \text{interv}_{Q_j}(\lambda)\text{”}, \quad (14)$$

$$\text{cost}_Q(\lambda) = \text{cost}_{Q_j}(\lambda), \quad (15)$$

$$\text{eff}_Q(\lambda) = \text{eff}_{Q_j}(\lambda). \quad (16)$$

The interpretation of this definition is as follows: for each value d_j (a possible choice) of the decision D there is CEP Q_j and for each value of λ there is an intervention $\text{interv}_{Q_j}(\lambda)$ in Q_j which is optimal for d_i . Equation 13 means that when making decision D we select j such that $\text{interv}_{Q_j}(\lambda)$ is the intervention that attains the highest NMB for that particular value of λ . Equation 14 means that the optimal intervention for decision D is to choose first the option d_j and then apply the intervention $\text{interv}_{Q_j}(\lambda)$. The cost and effectiveness associated with intervention $\text{interv}_Q(\lambda)$ —given by the optimal CEP, Q —are the same as in Q_j .

The key property of this definition is Equation 13, which states that for every λ the NMB of the optimal partition, Q , is the same as if we had performed a (unicriterion) maximization of the NMB for each single value of λ .

The optimal CEP can be obtained by applying Algorithm 3, which collects all the thresholds of the Q_j 's and performs a deterministic CEA (cf. Sec. 2.1) on each interval. Finally, it fuses some intervals by eliminating the unnecessary thresholds. In [1] we show with an example how this algorithm operates and why it is sometimes necessary to fuse intervals.

Algorithm 3: Optimal CEP.

Input: A set of m CEPs $\{Q_1, \dots, Q_m\}$, with $Q_j = (\Theta_j, C_j, E_j, I_j)$ and a decision node

Result: A new CEP $Q = (\Theta, C, E, I)$.

```

1  $\Theta \leftarrow \bigcup_{j=1}^m \theta_j$ 
2  $n \leftarrow \text{card}(\Theta)$ 
3 for  $i \leftarrow 1$  to  $n$  do
4    $\lfloor$  perform a deterministic CEA analysis of interval  $i$ 
5   fuse contiguous intervals having the same intervention, the same cost,
   and the same effectiveness

```

3.2 Construction of the ID

The construction of an ID for CEA is almost identical to the traditional case, in which the ID contains only one utility node because the decision is based on a single criterion. The difference is that in CEA we have two criteria, and consequently we put two utility nodes in the ID: U_c for the cost and U_e for the effectiveness. We will see an example below.

3.3 Evaluation of the influence diagram

Evaluation of the ID with the variable elimination algorithm Performing CEA with an ID is very similar to the evaluation of a (unicriterion) ID having only one decision node. The main differences are:

- In the unicriterion case, each potential assigns a real number to each configuration of its variables, while now we have a CEP-potential that assigns a CEP to each configuration.
- Initially, U is a CEP-potential that depends on all the variables that are parents of U_c or U_e . The CEP assigned to each configuration contains only one interval (no thresholds have been generated yet); the cost and the effectiveness are those obtained from the functions associated to U_c and U_e , which are kept separately in the CEP-potential.
- The weighted average and the maximization of potentials (lines 9 and 11 of the Algorithm 1) must be replaced by the weighted average of CEPs (Algorithm 2) and the computation of the optimal CEP (Algorithm 3).
- The potential returned by the algorithm is not a single real number, but a CEP.

The mathematical ground for this algorithm is as follows. If we know the value of λ , we can transform the cost-effectiveness problem into a one criterion problem by computing the NMB of each scenario using Equation 1. This way, the ID would have only one utility node instead of two and we might evaluate it with the Algorithm 1. As mentioned in the introduction, CEA is performed when we do not know the value of λ . The algorithm presented in this subsection is

equivalent to applying the Algorithm 1 for every single value of λ , but instead of doing an independent evaluation for each value of λ , which is clearly impossible, we group the λ 's into intervals having the same cost, the same effectiveness, and the same optimal intervention, and we evaluate all the values of λ in parallel. The CEP returned by the modified algorithm indicates, for each interval, the cost, the effectiveness, and the optimal intervention.

Evaluation of the ID with the arc reversal algorithm The basic idea of our method can be applied to the arc-reversal algorithm described in Section 2.2. The first step is to fuse the two utility nodes, U_c and U_e , into a single node U having an associated CEP-potential, the same as in the variable elimination algorithm for CEA. The operations of barren node removal and arc reversal are exactly the same as in the evaluation of unicriterion IDs.

The removal of a chance node is analogous to the case of unicriterion IDs, but now we have to perform a weighted average of CEPs (Algorithm 2) for each configuration of \mathbf{V} (cf. Eq. 6), because now $U_{\text{old}}(\mathbf{z})$ is a CEP-potential. Similarly, when removing a decision node we have to find the optimal CEP (Algorithm 3) for each configuration of \mathbf{V} (cf. Eq. 7).

4 Example: CEA of a test

Example 1. For a disease whose prevalence is 0.14, there are two possible therapies. The effectiveness of each therapy depends on whether the disease is present or not, as shown in Table 1. There is a test with a sensitivity of 90% and a specificity of 93%, and a cost of 150 €. Is the test cost-effective?

| Therapy | Cost | Effectiveness | |
|------------|---------|---------------|----------|
| | | +disease | ¬disease |
| No therapy | 0€ | 1.2 | 10.0 |
| Therapy 1 | 20,000€ | 4.0 | 9.9 |
| Therapy 2 | 70,000€ | 6.5 | 9.3 |

Table 1. Cost and effectiveness of each intervention for the Example 1.

This problem can be analyzed with the ID in Figure 2. The decision node *Dec:Test* represents the decision of performing the test or not, and *Therapy* represents the choice of therapy. The numerical information of the ID consists of four ordinary potentials (by “ordinary” we mean that they assign a real number to each configuration of their variables): $P(\text{disease})$, $P(\text{test}|\text{disease}, \text{dec:test})$, $U_e(\text{disease}, \text{therapy})$, and $U_c(\text{dec:test}, \text{therapy})$.

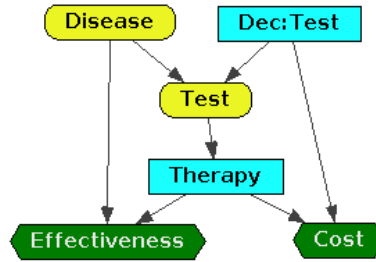


Fig. 2. Influence diagram for the Example 1.

Evaluation of this example with variable elimination The algorithm is initialized by building a CEP-potential U that depends on *Disease*, *Dec:Test*, and *Therapy*. For each of the $2 \times 2 \times 3 = 12$ configurations of these variables the CEP-potential U contains a CEP of only one interval, $(0, +\infty)$.

The first variable that the algorithm eliminates is *Disease*. The potential ψ is computed as the product of $P(\text{disease})$ and $P(\text{test}|\text{disease}, \text{dec:test})$. The potentials $\psi_{\text{ind}}(\text{dec:test}, \text{test})$ and $\psi_{\text{cond}}(\text{disease}, \text{dec:test}, \text{test}) = P(\text{disease}|\text{dec:test}, \text{test})$ are computed as indicated by the lines 6 and 7 in the Algorithm 1. The new CEP-potential U is

$$U(\text{dec:test}, \text{test}, \text{therapy}) = \sum_{\text{disease}} P(\text{disease}|\text{dec:test}, \text{test}) \cdot U(\text{disease}, \text{dec:test}, \text{therapy}) .$$

The elimination of the decision node *Therapy* is performed by applying a deterministic CEA (see Sec. 2.1) for each CEP in the new CEP-potential $U(\text{dec:test}, \text{test}, \text{therapy})$. This returns a new CEP-potential, $U(\text{dec:test}, \text{test})$.

The elimination of the chance node *Test* gives the following CEP-potential:

$$U(\text{dec:test}) = \sum_{\text{test}} \psi(\text{dec:test}, \text{test}) \cdot U(\text{dec:test}, \text{test}) .$$

Finally, the elimination of *Dec:Test* performs a deterministic CEA for each CEP in this CEP-potential, $U(\text{dec:test})$, which returns the CEP shown in Table 2.

Evaluation of this example with arc reversal The first step consists of fusing the two utility nodes, *Effectiveness* and *Cost*, into a single node U with the same associated CEP-potential as in the case of variable elimination: $U(\text{disease}, \text{dec:test}, \text{therapy})$.

The inversion of the arc $\text{Disease} \rightarrow \text{Test}$ leads to computing the potentials $P_{\text{new}}(\text{disease}|\text{test}, \text{dec:test})$ and $P_{\text{new}}(\text{test}|\text{dec:test})$. The first one is lost when eliminating the node *Disease*. The second is exactly the same potential $\psi(\text{dec:test}, \text{test})$ computed by the variable elimination algorithm. The consecutive elimination of

| Interval | Cost | Effectiveness | Dec:Test | Therapy |
|----------------------|--------|---------------|-------------|--|
| (0, 11,171) | 0 | 8.77 | Do not test | No therapy |
| (11,171, 33,384) | 3,874 | 9.11 | Do test | $\left\{ \begin{array}{l} \text{test:positive} \rightarrow \text{Therapy 1} \\ \text{test:negative} \rightarrow \text{No therapy} \end{array} \right.$ |
| (33,384, $+\infty$) | 13,184 | 9.39 | Do test | $\left\{ \begin{array}{l} \text{test:positive} \rightarrow \text{Therapy 2} \\ \text{test:negative} \rightarrow \text{No therapy} \end{array} \right.$ |

Table 2. Final CEP obtained by evaluating the influence diagram in Figure 2. It gives the cost, the effectiveness, and the optimal intervention for each value of λ .

Therapy, *Test*, and *Dec:Test*, which do not require any arc reversal, lead to the same utility potentials as in the case of variable elimination. In particular, the final CEP is the one shown in Table 2.

This example shows that even though the two algorithms look different, in fact they are performing essentially the same operations.

5 Related work

Nielsen et al. [11] have studied multi-attribute IDs, in which the global utility is given by

$$u = \alpha_1 \cdot u_1 + \dots + \alpha_n \cdot u_n \quad (17)$$

where each u_i is an attribute and the α_i 's represent the decision maker's preferences. The IDs presented in this paper are a particular case of the former, in which $n = 2$. However, the evaluation methods are completely different. The work by Nielsen et al. focuses on a particular configuration α , assumes that Δ_α^* , the optimal strategy for this α , is known (it is easy to evaluate a unicriterion ID), and tries to determine the support for this strategy, i.e., the region of \mathbb{R}^n for which the optimal strategy is Δ_α^* . The result of this analysis is a set of inequalities, which can be interpreted as the hyperplanes that delimit such region. A more intuitive way of summarizing the results of their analysis is to compute the radius of the biggest ball that can be contained in that region. In our opinion, this kind of analysis is far from the CEA that health decision makers demand.

In contrast, our study is limited to the case $n = 2$ and $\alpha_2 = -1$ (see Eq. 1), which allows us to find the optimal strategy as a function of α_1 , i.e., λ . We perform the kind of CEA that is usual in medicine, and the output of our algorithm can be summarized in the form of a table (see Table 2) whose interpretation is immediate: it shows the cost, the effectiveness, and the optimal intervention for each value of λ .

6 Conclusion and future work

In this paper we have presented two algorithms for performing CEA with influence diagrams (IDs); one of them is based on the variables elimination al-

gorithm [7]; the other is based on arc reversal [12, 14]. The main difference is that the standard algorithms operate with ordinary potentials, i.e., potentials that assign a real number to each configuration of their variables, while in our algorithms the conditional probabilities are ordinary potentials, but the utility is a CEP-potential, i.e., a potential that assigns a CEP to each configuration of its variables.

The algorithms presented in this paper are two adaptations of the method for performing CEA in decision trees with embedded decision nodes [1]. The main contribution of this paper is the possibility of solving IDs whose equivalent decision trees have prohibitive sizes. Using a Java implementation of the variable elimination algorithm, we have performed CEA on an ID for the mediastinal staging of non-small cell lung cancer that contains 5 decisions and 8 chance variables [9]. More recently, we have applied the same method to an ID for total knee replacement prosthesis that contains 4 decisions and 11 chance variables [8]. The equivalent decision trees, which can be obtained automatically from the influence diagrams, have thousands of leaves. Clearly, it would have been impossible to build those trees directly, and their evaluation would have been much less efficient than the evaluation of the IDs. An open line for future research is to summarize the results of these evaluations into small policy trees, using a method similar to the one proposed in [9].

Another line for future research is the adaptation of our CEA algorithms to the evaluation of decision analysis networks (DANs) [3], which present several advantages over IDs, especially in the case of asymmetric decision problems.

Acknowledgments

This work has been supported by grants TIN2006-11152 and TIN2009-09158, of the Spanish Ministry of Science and Technology, and by FONCICYT grant 85195.

References

1. M. Arias and F. J. Díez. Cost-effectiveness analysis with sequential decisions. Technical Report CISIAD-11-01, UNED, Madrid, Spain, 2011. <http://www.cisiad.uned.es/techreports/cea-multidec.php>.
2. C. Bielza, M. Gómez, and P. P. Shenoy. A review of representation issues and modelling challenges with influence diagrams. *Omega*, 39:227–241, 2011.
3. F. J. Díez and M. Luque. Representing decision problems with Decision Analysis Networks. Technical Report CISIAD-10-01, UNED, Madrid, Spain, 2010.
4. M. F. Drummond, M. J. Sculpher, G. W. Torrance, B. J. O’Brien, and G. L. Stoddart. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, third edition, 2005.
5. M. R. Gold, J. E. Siegel, L. B. Russell, and M. C. Weinstein. *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York, 1996.
6. R. A. Howard and J. E. Matheson. Influence diagrams. In R. A. Howard and J. E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis*, pages 719–762. Strategic Decisions Group, Menlo Park, CA, 1984.

7. F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York, second edition, 2007.
8. D. León. An influence diagram for total knee arthroplasty. Master's thesis, Dept. Artificial UNED, Madrid, Spain, 2011.
9. M. Luque. *Probabilistic Graphical Models for Decision Making in Medicine*. PhD thesis, UNED, Madrid, 2009.
10. M. Luque and F. J. Díez. Variable elimination for influence diagrams with super-value nodes. *International Journal of Approximate Reasoning*, 51(6):615 – 631, 2010.
11. S. H. Nielsen, T. D. Nielsen, and F. V. Jensen. Multi-currency influence diagrams. In A. Salmerón and J. A. Gámez, editors, *Advances in Probabilistic Graphical Models*, pages 275–294. Springer, Berlin, Germany, 2007.
12. S. M. Olmsted. *On Representing and Solving Decision Problems*. PhD thesis, Dept. Engineering-Economic Systems, Stanford University, CA, 1983.
13. H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. MIT press, Cambridge, 1961.
14. R. D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.
15. H. C. Sox, M. A. Blatt, M. C. Higgins, and K. I. Marton. *Medical Decision Making*. Butterworth-Heinemann, Woburn, MA, 1988.
16. A. A. Stinnett and J. Mullahy. Net health benefit: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18:S68–S80, 1998.
17. J. A. Tatman and R. D. Shachter. Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man, and Cybernetics*, 20:365–379, 1990.
18. M. C. Weinstein and W. B. Stason. Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine*, 296:716–721, 1977.
19. M. C. Weinstein, G. Torrance, and A. McGuire. QALYs: The basics. *Value in Health*, 12(Supplement 1):S5–S9, 2009.

Impact of Quality of Bayesian Network Parameters on Accuracy of Medical Diagnostic Systems

Agnieszka Onisko^{1,3} and Marek J. Druzdziel^{1,2}

¹ Faculty of Computer Science, Białystok University of Technology, Wiejska 45A, 15-351 Białystok, Poland

² Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Programs, University of Pittsburgh, Pittsburgh, PA 15260, USA

³ Magee-Womens Hospital, Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA

Abstract. While most knowledge engineers believe that the quality of results obtained by means of Bayesian networks is not too sensitive to imprecision in probabilities, this remains a conjecture with only modest empirical support. We summarize the results of several previously presented experiments involving HEPAR II model, in which we manipulated the quality of the model's numerical parameters and checked the impact of these manipulations on the model's accuracy. The chief contribution of this paper are results of replicating our experiments on several medical diagnostic models derived from data sets available at the Irvine Machine Learning Repository. We show that the results of our experiments are qualitatively identical to those obtained earlier with HEPAR II.

1 Introduction

Decision-analytic methods provide a coherent framework for modeling and solving decision problems in decision support systems [12]. A valuable modeling tool for complex uncertain domains, such as those encountered in medical applications, is a Bayesian network [19], an acyclic directed graph quantified by numerical parameters and modeling the structure of a domain and the joint probability distribution over its variables. There exist algorithms for reasoning in Bayesian networks that compute the posterior probability distribution over some variables of interest given a set of observations. As these algorithms are mathematically correct, the ultimate quality of their results depends directly on the quality of the underlying models and their parameters. These parameters are rarely precise, as they are often based on subjective estimates or data that do not reflect precisely the target population.

The question of sensitivity of Bayesian networks to precision of their parameters is of much interest to builders of intelligent systems. If precision does not matter, rough estimates or even qualitative "order of magnitude" estimates that are typically obtained in the early phases of model building, should be sufficient

without the need for their painstaking refinement. Conversely, if network results are sensitive to the precise values of probabilities, a lot of effort has to be devoted to obtaining precise estimates.

There is a popular belief, supported by anecdotal evidence, that Bayesian network models are tolerant to imprecision in their numerical parameters. Pradhan *et al.* [20] were the first to describe an experiment in which they studied the behavior of a large medical diagnostic model, the CPCS network [15, 23]. Their key experiment, which we will subsequently refer to as the *noise propagation experiment*, focused on systematic introduction of noise in the original parameters (assumed to be the gold standard) and measuring the influence of the amount of noise on the average posterior probability of the true diagnosis. They observed that this average was insensitive to even very large amounts of noise. The noise propagation experiment, while ingenious and thought provoking, offers room for improvements. The first problem, pointed out by Coupé and van der Gaag [7], is that the experiment focused on the average posterior rather than individual posterior in each diagnostic case and how it varies with noise, which is of most interest. The second weakness is that the posterior of the correct diagnosis is by itself not a sufficient measure of model robustness. Practical model performance will depend on how these posteriors are used. In order to make a rational diagnostic decision, for example, one needs to know at least the probabilities of rival hypotheses (and typically the joint probability distribution over all disorders). Only this allows for weighting the utility of correct against the dis-utility of incorrect diagnosis. If the focus of reasoning is differential diagnosis, it is of importance to observe how the posterior in question compares to the posteriors of competing disorders. Another problem is that noise introduced in parameters was assumed to be random, which may not be a reasonable assumption. It is known, for example, that human experts often tend to be overconfident [16]. Yet another opportunity for improvement is looking at precision of parameters rather than their random deviations from the true value. Effectively, the results of the noise propagation experiment are tentative and the question whether actual performance of Bayesian network models is robust to imprecision in their numerical parameters remains open.

Search for those parameters whose values are critical for the overall quality of decisions is known as sensitivity analysis. Sensitivity analysis studies how much a model output changes as various model parameters vary through the range of their plausible values. It allows to get insight into the nature of the problem and its formalization, helps in refining the model so that it is simple and elegant (containing only those factors that matter), and checks the need for precision in refining the numbers [16]. Several researchers proposed efficient algorithms for performing sensitivity analysis in Bayesian networks (e.g., [3, 6, 7, 14]). It is theoretically possible that small variations in a numerical parameter cause large variations in the posterior probability of interest. Van der Gaag and Renooij [11] found that practical networks may indeed contain such parameters. Because practical networks are often constructed with only rough estimates of probabilities, a question of practical importance is whether overall imprecision

in network parameters is important. If not, the effort that goes into polishing network parameters might not be justified, unless it focuses on their small subset that is shown to be critical.

In this paper, we report the results of a series of experiments in which we manipulate the quality of parameters of several real or realistic Bayesian network models and study the impact of this manipulation on the precision of their results. In addition to looking at symmetric noise, like in the original noise propagation experiment, we enter noise in the parameters in such a way that the resulting distributions become biased toward extreme probabilities, hence, modeling expert overconfidence in probability estimates. Our results show that the diagnostic accuracy of Bayesian network models is sensitive to imprecision in probabilities. It appears, however, that it is less sensitive to overconfidence in probabilities than it is to symmetric noise. We also test the sensitivity of models to underconfidence in parameters and show that underconfidence in parameters leads to more error than symmetric noise.

We examine also a related question: “Are Bayesian networks sensitive to precision of their parameters?” Rather than entering noise into the parameters, we change their precision, starting with the original values and rounding them systematically to progressively rougher scales. This models a varying degree of precision of the parameters. Our results show that the diagnostic accuracy of Bayesian networks is sensitive to imprecision in probabilities, if these are plainly rounded. However, the main source of this sensitivity appears to be in rounding small probabilities to zero. When zeros introduced by rounding are replaced by very small non-zero values, imprecision resulting from rounding has minimal impact on diagnostic performance.

Our experiments suggest that Bayesian networks may be less sensitive to the quality of their numerical parameters than previously believed. While noise in numerical parameters starts taking its toll almost from the very beginning, there is a noticeable region of tolerance to small amounts of noise.

The remainder of this paper is structured as follows. Section 2 introduces the models used in our experiments. Section 3 describes our experiments based on introducing noise into probabilities. Section 4 describes our experiments based on progressive rounding of parameters. Finally, Section 5 summarizes our results and main insights obtained from these results.

2 Models studied

The main model used in our experiments is the HEPAR II model [18]. This is one of the largest practical medical Bayesian network models available to the community, carefully developed in collaboration with medical experts and parametrized using clinical data.⁴ We would like to note that the results for the HEPAR II network presented in this paper have been presented before [9, 10, 17]. In addition, we selected three data sets from the Irvine Machine Learning Repository:

⁴ Readers interested in HEPAR II can download it from Decision Systems Laboratory’s model repository at <http://genie.sis.pitt.edu/>.

Table 1. Medical data used in our experiments

| data set | instances | variables | variable types | classes |
|--------------------|-----------|-----------|----------------------|---------|
| Acute Inflammation | 120 | 8 | categorical, integer | 4 |
| SPECT Heart | 267 | 22 | categorical | 2 |
| Cardiotocography | 2,126 | 23 | categorical, real | 3 |
| HEPAR II | 699 | 70 | categorical, real | 11 |

(1) Acute inflammation [8], (2) SPECT Heart [4], and (3) Cardiotocography [22]. Table 1 presents basic characteristics of the selected data sets, including HEPAR data. Table 2 presents basic statistics of Bayesian network models that we created from the data. All models consist of only discrete nodes with all continuous variables discretized before the models were learned.

Table 2. Bayesian network models used in our experiments

| model | nodes | arcs | states | parameters | avg in-degree | avg outcomes |
|--------------------|-------|------|--------|------------|---------------|--------------|
| ACUTE INFLAMMATION | 8 | 15 | 17 | 97 | 1.88 | 2.13 |
| SPECT HEART | 23 | 52 | 46 | 290 | 2.26 | 2.00 |
| CARDIOTOCOGRAPHY | 22 | 63 | 64 | 13,347 | 2.86 | 2.91 |
| HEPAR II | 70 | 121 | 162 | 2,139 | 1.73 | 2.24 |

Similarly to Pradhan *et al.* [20], for the purpose of our experiments, we assumed that the model parameters were perfectly accurate and, effectively, the diagnostic performance achieved was the best possible. Of course, in reality, the parameters of the model may not be accurate and the performance of the model can be improved upon. In our experiments, we study how this baseline performance degrades under the condition of noise and inaccuracy.

We define diagnostic accuracy as the percentage of correct diagnoses on real patient cases. This is obviously a simplification, as one might want to know the sensitivity and specificity data for each of the disorder or look at the global quality of the model in terms of AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristics) curve, as suggested by a reviewer. This, however, is complicated in case of models focusing on multiple disorders — there is no single measure of performance but rather a measure of performance for every single disorder. We decided thus to focus on the percentage of correct diagnoses.

Because Bayesian network models operate only on probabilities, we assume that each model indicates as correct the diagnosis that is most likely given evidence. When testing the accuracy of models, we were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of w most probable diagnoses contains the correct diagnosis for small values of w (we chose a “window” of $w=1, 2, 3$,

and 4). The latter focus is of interest in diagnostic settings, where a decision support system only suggest possible diagnoses to a physician. The physician, who is the ultimate decision maker, may want to see several top alternative diagnoses before focusing on one.

3 Noise in parameters

Our first series of experiments focused on sensitivity of accuracy of Bayesian network models to symmetric noise in their parameters. When introducing noise into model parameters, we used the approach proposed by Pradhan *et al.* [20], which is transforming each original probability into log-odds form, adding symmetric Gaussian noise parametrized by a parameter σ , and transforming it back to probability, i.e.,

$$p' = Lo^{-1}[Lo(p) + \text{Normal}(0, \sigma)] , \quad (1)$$

where

$$Lo(p) = \log_{10}[p/(1 - p)] . \quad (2)$$

This guarantees that the transformed probability lies within the interval $(0, 1)$.

3.1 Symmetric noise

In [17], we performed experiments focusing on how symmetric noise (see the top two graphs in Figure 2 to get an idea of what this noise amounts to) introduced into network parameters affects the diagnostic accuracy of HEPAR II. Figure 1 presents the diagnostic accuracy of 30 versions of the network (each for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on the set of test cases for different values of window size as a function of σ .

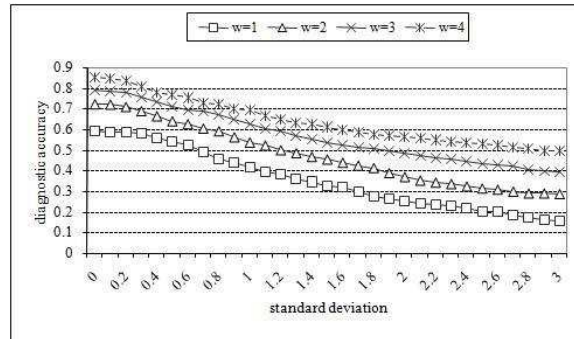


Fig. 1. The diagnostic accuracy of the model under symmetric noise as a function of σ ($w=1$) [17].

Diagnostic performance seems to deteriorate for even smallest values of noise, although it has to be said that the plot shows a small region (for σ smaller than roughly 0.2) in which performance loss is minimal.

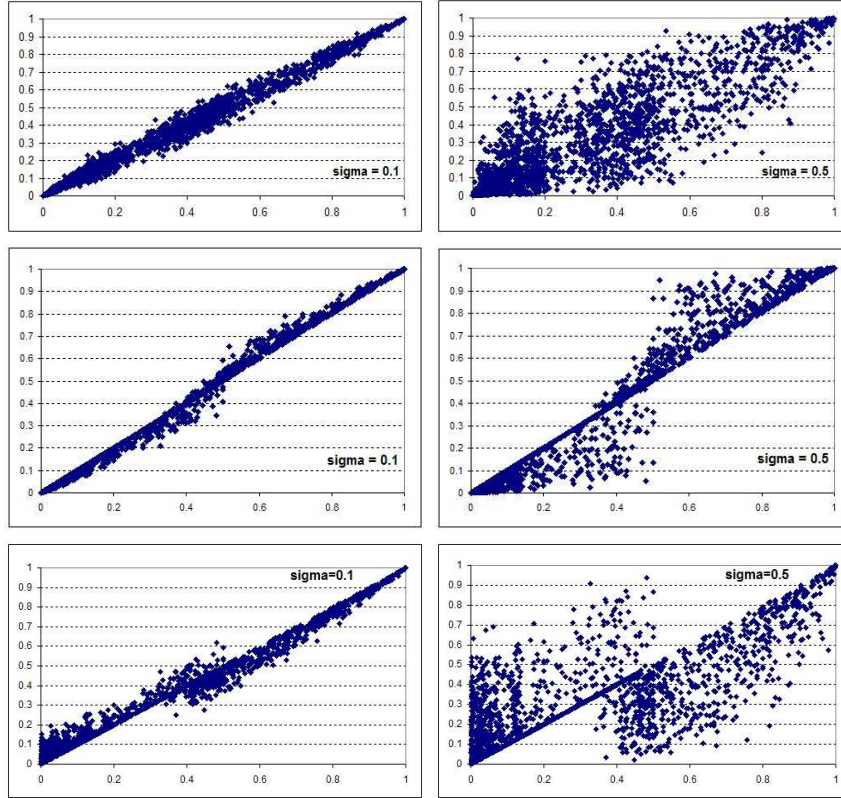


Fig. 2. Scatterplots of the original (horizontal axis) vs. transformed (vertical axis) probabilities for $\sigma = 0.1$ and $\sigma = 0.5$. The top two plots show symmetric noise, the middle two plots show overconfidence, the bottom two plots show underconfidence.

3.2 Biased noise

Symmetric random noise does not seem to be very realistic. It is a known tendency of experts to be overconfident about their probability estimates, i.e., offer more extreme probability estimates than warranted by objective evidence [13, 16]. One way of simulating bias in expert judgment is to distort the original parameters so that they become more extreme (this amounts to modeling expert overconfidence) or more centered, i.e., biased towards uniform probabilities

(this amounts to modeling expert underconfidence). Our next experiment (reported in [9]) focused on investigating the influence of biased noise in HEPAR II's probabilities on its diagnostic performance.

We introduced bias into noise in the following way. Given a discrete probability distribution Pr , for overconfidence, we identified the smallest probability p_S . We transformed this smallest probability p_S into p'_S by making it even smaller, according to the following formula:

$$p'_S = Lo^{-1}[Lo(p_S) - |\text{Normal}(0, \sigma)|] .$$

We made the largest probability in the probability distribution Pr , p_L , larger by precisely the amount by which we decreased p_S , i.e.,

$$p'_L = p_L + p_S - p'_S .$$

An alternative way of introducing biased noise suggested to us is by means of building a logistic regression/IRT model (e.g., [1, 2, 21]) for each conditional probability table and, subsequently, manipulating the slope parameter. For underconfidence, we identified the highest probability p_L . We then transformed p_L into p'_L by making it smaller, according to the following formula:

$$p'_L = Lo^{-1}[Lo(p_L) - |\text{Normal}(0, \sigma)|] .$$

We made the smallest probability in the probability distribution Pr , p_S , higher by precisely the amount by which we decreased p_L , i.e.,

$$p'_S = p_S + p_L - p'_L .$$

We were by this guaranteed that the transformed parameters of the probability distribution Pr' added up to 1.0.

Figure 2 shows the effect of introducing this biased noise. The middle two plots in the figure show overconfidence transformation and the bottom two show underconfidence. For overconfidence, in particular, the transformation is such that small probabilities are likely to become smaller and large probabilities are likely to become larger. Effectively, the distributions become more biased towards extreme probabilities.

We tested 30 versions of HEPAR II for each of the conditions (each network for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on all records of the HEPAR data set and computed HEPAR II's diagnostic accuracy. We plotted this accuracy in Figure 3 as a function of σ for different values of window size w . The left plot is for the overconfidence and the right plot is for the underconfidence condition.

It is clear that HEPAR II's diagnostic performance deteriorates with biased noise as well. The results are qualitatively similar to those in Figure 1, although performance under overconfidence bias degraded more slowly with the amount of noise than performance under symmetric noise, which, in turn degraded more slowly than performance under underconfidence.

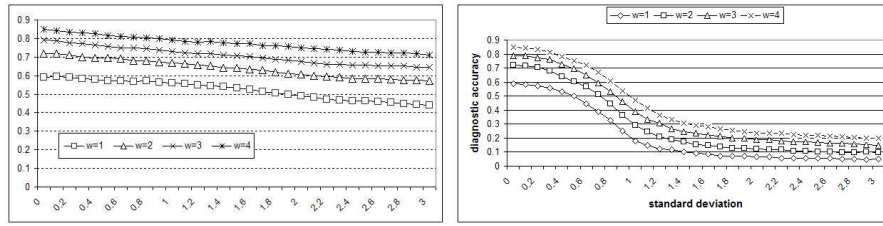


Fig. 3. The diagnostic accuracy of HEPAR II for various window sizes as a function of the amount of biased noise (expressed by σ). Overconfidence (left plot) and underconfidence (right plot) [10].

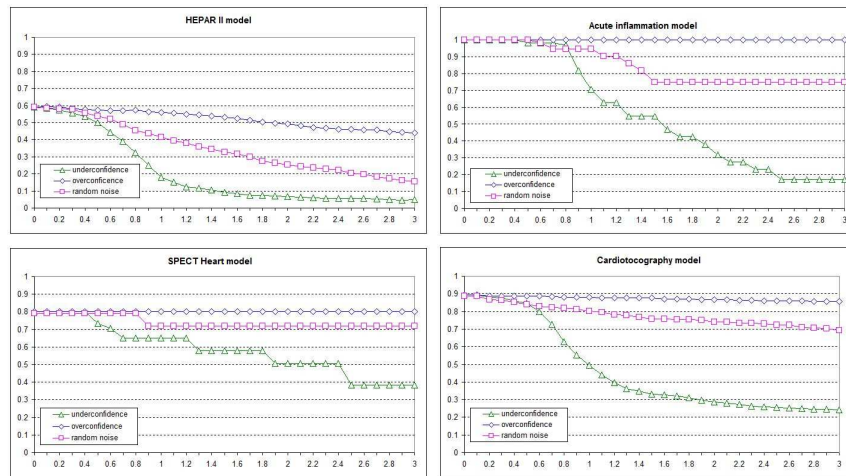


Fig. 4. The diagnostic accuracy of the four models (clock-wise HEPAR II, ACUTE INFLAMMATION, SPECT HEART and CARDIOTOGRAPHY) as a function of the amount of biased and unbiased noise, window $w = 1$.

We repeated this experiment for the three networks from the Irvine repository. Figure 4 shows the accuracy of HEPAR II and the three Irvine models as a function of the amount of biased and unbiased noise, window $w = 1$, on the same plot. The results are qualitatively identical: performance under underconfidence bias in all four cases degrades faster than performance under symmetric and overconfident noise.

It is interesting to note that here again for small values of σ , there is only a minimal effect of noise on performance.

4 Imprecision in parameters

Our next step was investigating how progressive rounding of a Bayesian network's probabilities affects its diagnostic performance. To that effect, we have

successively created various versions of models with different precision of parameters and tested the performance of these models.

For the purpose of our experiment, we used $n = 100, 10, 5, 4, 3, 2$, and 1 , for the number of intervals in which the probabilities fall. And so, for $n = 10$, we divided the probability space into 10 intervals and each probability took one of 11 values, i.e., $0.0, 0.1, 0.2, \dots, 0.9$, and 1.0 . For $n = 5$, each probability took one of six values, i.e., $0.0, 0.2, 0.4, 0.6, 0.8$, and 1.0 . For $n = 2$, each probability took one of only three values, i.e., $0.0, 0.5$, and 1.0 . Finally, for $n = 1$, the smallest possible value of n , each probability was either 0.0 or 1.0 . Figure 5 shows scatter plots of all 2,139 HEPAR II's parameters (horizontal axis) against their rounded values (vertical axis) for n equal to 10, 5, 2, and 1.

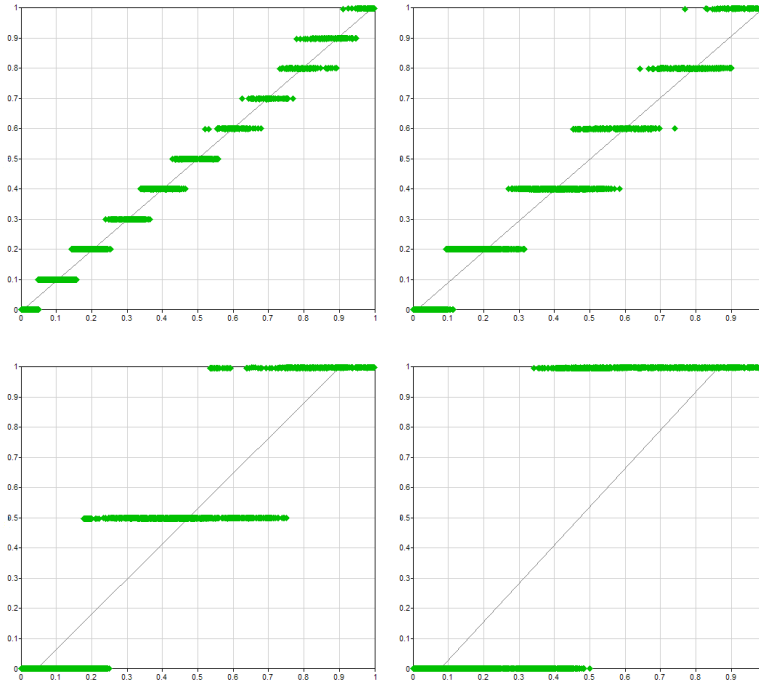


Fig. 5. Rounded vs. original probabilities for various levels of rounding accuracy.

Please note the drastic reduction in precision of the rounded probabilities, as pictured by the vertical axis. When $n = 1$, all rounded probabilities are either 0 or 1. Also, note that the horizontal bars in the scatter plot overlap. For example, in the upper-right plot ($n = 5$), we can see that an original probability $p = 0.5$ in HEPAR II got rounded sometimes to 0.4 and sometimes to 0.6. This is a simple consequence of the surrounding probabilities in the same distribution and the

necessity to make the sum of rounded probabilities add to 1.0, as guaranteed by the algorithm that we used for rounding probabilities.

We computed the diagnostic accuracy of various versions of HEPAR II, as produced by the rounding procedure. Figure 6 shows a summary of the results in both graphical and tabular format. The horizontal axis in the plot corresponds to the number of intervals n in logarithmic scale, i.e., value 2.0 corresponds to the rounding $n = 100$, and value 0 to the rounding $n = 1$. Intermediate points, for the other roundings can be identified in-between these extremes. The numerical accuracy reported in the table corresponds to the lower curve in the plot.

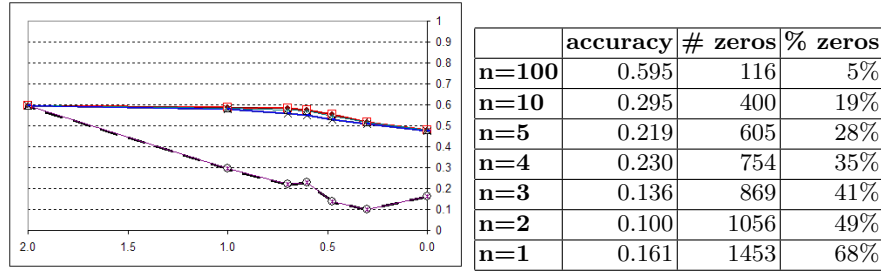


Fig. 6. Diagnostic performance of HEPAR II as a function of logarithm of parameter accuracy and ε ($w=1$) [9].

It turns out that the strongly deteriorating accuracy is the effect of zeros in the probability distributions introduced by rounding. Please note that zero in probability theory is a special value. Once the probability of an event becomes zero, it can never change, no matter how strong the evidence for it. We addressed this problem by replacing all zeros introduced by the rounding algorithm by small ε probabilities and subtracting the introduced ε s from the probabilities of the most likely outcomes in order to preserve the constraint that the sum should be equal to 1.0. While this caused a small distortion in the probability distributions (e.g., a value of 0.997 instead of 1.0 when $\varepsilon = 0.001$ and there were three induced zeros transformed into ε), it did not introduce sufficient difference to invalidate the precision loss. To give the reader an idea of what it entailed in practice, we will reveal the so far hidden information that the plots in Figure 5 were obtained for data with $\varepsilon = 0.001$.

The result of this modification was dramatic and is pictured by the upper curves in Figure 6, each line for a different value of ε . As can be seen, the actual value of ε did not matter too much (we tried three values: 0.0001, 0.001, and 0.01). In each case HEPAR II's performance was barely affected by rounding, even when there was just one interval, i.e., when all probabilities were either ε or $1 - \varepsilon$.

Our next experiment focused on the influence of precision in probabilities on HEPAR II's accuracy for windows of size 1, 2, 3, and 4. Figure 7 shows a summary of the results in both graphical and tabular format. The meaning of

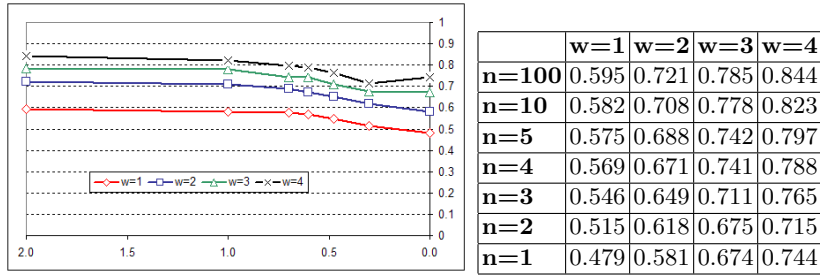


Fig. 7. Diagnostic performance of HEPAR II as a function of the logarithm of parameter accuracy and various window sizes [9].

the horizontal and vertical axes is the same as in Figure 6. We can see that the stability of HEPAR II's performance is similar for all window sizes.

We repeated the rounding experiment for the three networks from the Irvine repository. Figure 8 shows the accuracy of HEPAR II and the three Irvine models (window $w = 1$) as a function of the logarithm of parameter accuracy on the same plot. The results were qualitatively identical to those involving HEPAR II.

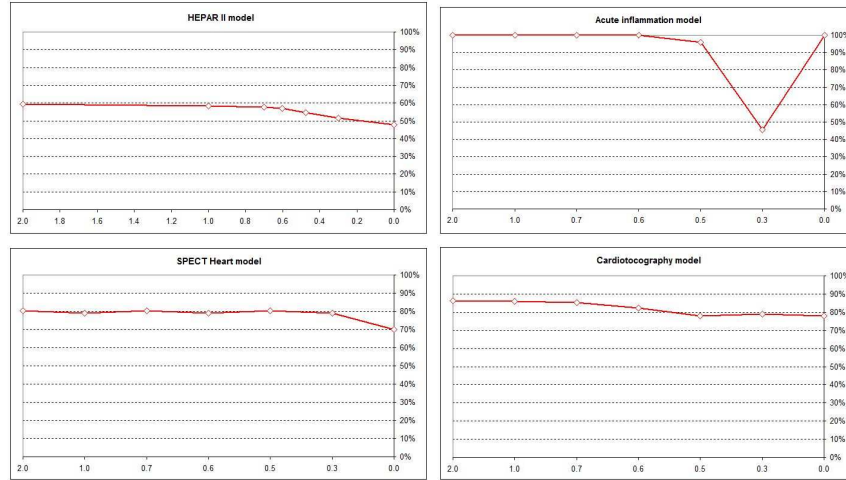


Fig. 8. The diagnostic accuracy of the four models (clock-wise HEPAR II, ACUTE INFLAMMATION, SPECT HEART and CARDIOTOGRAPHY) as a function of the logarithm of parameter accuracy, window $w = 1$.

5 Discussion

We described a series of experiments studying the influence of precision in parameters on model performance in the context of a practical medical diagnostic model, HEPAR II (these results were previously published in [9, 10, 17]), and three additional models based on real medical data from the Irvine Machine Learning Repository. We believe that the study was realistic in the sense of studying real models and focusing on a practical performance measure.

Our study has shown that the performance of all four models is sensitive to noise in numerical parameters, i.e., the diagnostic accuracy of the models decreases after introducing noise into their numerical parameters. For small to moderate amounts of noise, i.e., σ smaller than say 0.2, the effect of noise on accuracy was minimal. The effect of rounding the parameters was also minimal, giving some support to insensitivity of Bayesian network models to precision of their parameters.

We studied the influence of bias in parameters on model performance. Overconfidence bias had in our experiments a smaller negative effect on model performance than random noise. Underconfidence bias led to most serious deterioration of performance. While it is only a wild speculation that begs for further investigation, one might see our results as an explanation why humans tend to be overconfident rather than underconfident in their probability estimates. An interesting suggestion on the part of one of the reviewers was the link between bias, as we formulated it, and entropy. Models with parameters biased toward underconfidence have higher entropy and, thus, contain less information than models with symmetric noise or models biased toward overconfidence.

Our study of the influence of precision in parameters on model performance was inspired by the work of Clancey and Cooper [5], who conducted an experiment probing the sensitivity of MYCIN to the accuracy of its numerical specifications of degree of belief, certainty factors (CF). They applied a progressive roughening of CFs by mapping their original values onto a progressively coarser scale. The CF scale in MYCIN had 1,000 intervals ranging between 0 and 1,000. If this number was reduced to two, for example, every positive CF was replaced by the closest of the following three numbers: 0, 500, and 1,000. Roughening CFs to hundred, ten, five, three, and two intervals showed that MYCIN is fairly insensitive to their accuracy. Only when the number of intervals was reduced to three and two, there was a noticeable effect on the system performance.

Our results are somewhat different. It appears that the diagnostic accuracy of Bayesian network models is sensitive to imprecision in probabilities, if these are rounded. However, the main source of this sensitivity appears to be in rounding small probabilities to zero. When zeros introduced by rounding are replaced by very small non-zero values, imprecision resulting from rounding has minimal impact on Bayesian network model's performance.

Acknowledgments

Agnieszka Onisko was supported by the Białystok University of Technology grants W/WI/1/02 and S/WI/2/2008, by the MNiI (Ministerstwo Nauki i Informatyzacji) grant 3T10C03529, and by the Polish Committee for Scientific Research grant 4T11E05522. Marek Druzdzel was supported by the National Institute of Health under grant number U01HL101066-01, by the Air Force Office of Scientific Research grants F49620-00-1-0112, F49620-03-1-0187, and FA9550-06-1-0243, and by Intel Research.

While we are solely responsible for any remaining shortcomings of this paper, our work has benefitted from helpful comments and suggestions from several individuals, of whom we would like to thank in particular Greg Cooper and Linda van der Gaag. Anonymous reviewers asked several excellent questions and offered suggestions that led to improvements of the paper.

All Bayesian network models in this paper were created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory and available at <http://genie.sis.pitt.edu/>.

References

1. Almond, R.G., Dibello, L.V., Jenkins, F., Mislevy, R., Senturk, D., Steinberg, L., Yan, D.: Models for conditional probability tables in educational assessment. In: Jaakkola, T., Richardson, T. (eds.) *Artificial Intelligence and Statistics*. pp. 137–143. Morgan Kaufmann (2001)
2. Almond, R.G., Dibello, L.V., Moulder, B., Zapata-Rivera, J.D.: Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement* 44(4), 341–359 (2007)
3. Chan, H., Darwiche, A.: When do numbers really matter? *Journal of Artificial Intelligence Research* 17, 265–287 (2002)
4. Cios, K.J., Kurgan, L.A.: UCI machine learning repository (2011), <http://archive.ics.uci.edu/ml>
5. Clancey, W.J., Cooper, G.: Uncertainty and evidential support. In: Buchanan, B.G., Shortliffe, E.H. (eds.) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chap. 10, pp. 209–232. Addison-Wesley, Reading, MA (1984)
6. Coupé, V.H.M., van der Gaag, L.: Practicable sensitivity analysis of Bayesian belief networks. In: *Prague Stochastics '98 — Proceedings of the Joint Session of the 6th Prague Symposium of Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*. pp. 81–86. Union of Czech Mathematicians and Physicists (1998)
7. Coupé, V.H.M., van der Gaag, L.C.: Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence* 36, 323–356 (2002)
8. Czerniak, J., Zarzycki, H.: Application of rough sets in the presumptive diagnosis of urinary system diseases. In: *Artificial Intelligence and Security in Computing Systems, ACS'2002 9th International Conference*. pp. 41–51. Kluwer Academic Publishers (2003)

9. Druzdzetel, M.J., Onisko, A.: Are Bayesian networks sensitive to precision of their parameters? In: S.T. Wierzchoń, M.K., Michalewicz, M. (eds.) *Proceedings of the Intelligent Information Systems Conference (XVI)*. pp. 35–44. Academic Publishing House EXIT, Warsaw, Poland (2008)
10. Druzdzetel, M.J., Onisko, A.: The impact of overconfidence bias on practical accuracy of Bayesian network models: An empirical study. In: *Working Notes of the 2008 Bayesian Modelling Applications Workshop, Special Theme: How Biased Are Our Numbers? Part of the Annual Conference on Uncertainty in Artificial Intelligence (UAI-2008)*. Helsinki, Finland (2008)
11. van der Gaag, L.C., Renooij, S.: Analysing sensitivity data from probabilistic networks. In: *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2001)*. pp. 530–537. Morgan Kaufmann Publishers, San Francisco, CA (2001)
12. Henrion, M., Breese, J.S., Horvitz, E.J.: *Decision Analysis and Expert Systems*. *AI Magazine* 12(4), 64–91 (Winter 1991)
13. Kahneman, D., Slovic, P., Tversky, A. (eds.): *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge (1982)
14. Kjaerulff, U., van der Gaag, L.C.: Making sensitivity analysis computationally efficient. In: *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*. pp. 317–325. Morgan Kaufmann Publishers, San Francisco, CA (2000)
15. Middleton, B., Shwe, M., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., Cooper, G.: Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: II. Evaluation of diagnostic performance. *Methods of Information in Medicine* 30(4), 256–267 (1991)
16. Morgan, M.G., Henrion, M.: *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge (1990)
17. Onisko, A., Druzdzetel, M.J.: Effect of imprecision in probabilities on Bayesian network models: An empirical study. In: *Working notes of the European Conference on Artificial Intelligence in Medicine (AIME-03): Qualitative and Model-based Reasoning in Biomedicine*. Protaras, Cyprus (October 18–22 2003)
18. Onisko, A., Druzdzetel, M.J., Wasyluk, H.: Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning* 27(2), 165–182 (2001)
19. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA (1988)
20. Pradhan, M., Henrion, M., Provan, G., del Favero, B., Huang, K.: The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence* 85(1–2), 363–397 (Aug 1996)
21. Rijmen, F.: Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning* In press
22. Marques de Sá, J., Bernardes, J., Ayres de Campos, D.: *UCI Machine Learning Repository* (2011), <http://archive.ics.uci.edu/ml>
23. Shwe, M., Middleton, B., Heckerman, D., Henrion, M., Horvitz, E., Lehmann, H., Cooper, G.: Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: I. The probabilistic model and inference algorithms. *Methods of Information in Medicine* 30(4), 241–255 (1991)

Author Index

- Ávila-Rios, Santiago 41
- Angelopoulos, Nicos 69
- Antal, P. 55
- Arias, Manuel 93
- Avilés, Héctor 14
- Bielza, Concha 29
- Borchani, Hanen 29
- Díez, Francisco Javier 93
- Davis, Jesse 67
- Demetrovics, Zs. 55
- Druzdzal, Marek J. 109
- Fenton, Norman 79
- González, Jesus A. 41
- Hauskrecht, Milos 1
- Heijden, Maarten van der 69
- Hernández-Franco, Jorge 14
- Hernandez-Leal, Pablo 41
- Hommersom, Arjen 79
- Lappenschaar, Martijn 79
- Larrañaga, Pedro 29
- Leder, Ronald 14
- Lucas, Peter J.F. 69, 79
- Luis, Roger 14
- Marsch, William 79
- Marx, P. 55
- Morales, Eduardo F. 41
- Nagarajan, Radhakrishnan 15
- Nemoda, Zs. 55
- Onisko, Agnieszka 109
- Orihuela-Espina, Felipe 14, 41
- Oropeza, Juan 14
- Perkins, Zane 79
- Reyes-Terán, Gustavo 41
- Rios-Flores, Alma 41
- Sarkozy, P. 55
- Sasvari-Szekely, M. 55
- Scutari, Marco 15
- Sucar, L. Enrique 14, 41
- Szekely, A. 55
- Varga, G. 55
- Visscher, Stefan 79
- Wessels, Lodewyk 69
- Yet, Barbaros 79